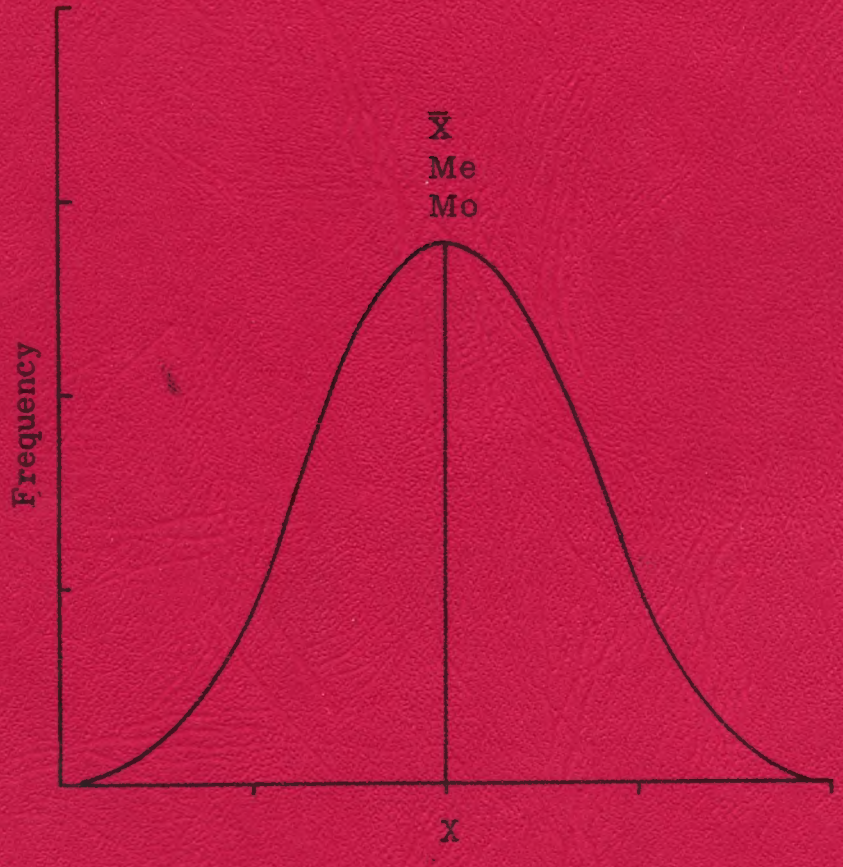
STATISTICAL TECHNIQUES IN FORESTRY

Andrew J. Nash, Ph.D.
Professor of Forestry
University of Missouri



Columbia, Mo.
Los Angeles

Lucas Brothers Publishers

STATISTICAL TECHNIQUES
IN
FORESTRY

Andrew J. Nash, Ph.D.
Professor of Forestry
University of Missouri

Lucas Brothers Publishers
909 LOWRY
COLUMBIA, MISSOURI 65201

2nd Edition

Printed in the United States of America

TABLE OF CONTENTS

PREFACE

Foresters have a definite need for some statistical training regardless of the general or specialized field in which they might be employed. The need is evident. The question confronting teachers of statistics is: what is the minimum course content necessary to give adequate training? For the majority of students in forestry education, an introductory course in statistics will be their only exposure to the subject. For those who have a definite goal of graduate study leading to research, an introductory course is not sufficient. These students will undoubtedly take more advanced work in statistical analysis covering such subjects as covariance analysis, advanced experimental design, multivariate analysis and so on.

For the forestry student, a course in statistics is either required in the curriculum or is recommended as an elective subject. To be effective, a course in statistical techniques should be adequate enough to give him an appreciation of the subject; the student should come to realize what he can do with statistics. More importantly, he should also realize what he CAN NOT do in the name of statistics.

Minor changes have been made in this edition of STATISTICAL TECHNIQUES IN FORESTRY. Additions have been made to the chapter on sampling to include a better presentation of unrestricted random sampling and to include a section on cluster sampling.

Problems appearing in the Appendix may be assigned by the instructor or students may do these problems on their own. Familiarity with statistical problems, rather than breeding contempt, makes one appreciate the role of statistical techniques in the forestry profession. Familiarity comes through solving numerous problems.

This text fulfills the promise made in the foreword to the 1st. edition of "Elementary Statistics for Foresters" for the inclusion of more advanced work in statistical techniques applied to the forestry profession. It is the author's hope that teachers and students will benefit by this treatment of the subject matter.

Grateful acknowledgement is made to my colleagues who not only gave me encouragement in the preparation of this text but also supplied some data which have been most helpful. My thanks are given to George W. Snedecor and his publishers, the Iowa State University Press for permission to reproduce Tables 5.5, A.1 and A.4 and parts of Tables 6.1 and 9.3. I am indebted to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., Cambridge, and to Oliver and Boyd Ltd., Edinburgh, for their permission to reprint Table IV from their book "Statistical Methods for Research Workers". I am also grateful to Professor E.S. Pearson who, on behalf of the Trustees of Biometrika, granted permission to reproduce parts of Table 6.1 (the t-distribution) and Table 9.3, the 5% points of Q.

Columbia, Missouri

Andrew J. Nash

INTRODUCTION

In addition to a knowledge of the methods of measuring trees, the computation of their volumes, the construction of yield tables and the collection of field data, a grasp of statistical methods is essential to the practicing forester.

In any profession or business, certain tools of the trade are necessary. A mechanic has a tool box in which are wrenches, pliers, micrometer gauges and a various assortment of other specialized tools. In the forester's "tool box" will be found, among many varied and assorted facts and theories, a familiarity with basic statistical techniques. With these tools, he is able to make more intelligent judgments, to ascertain the probability of events and to recommend procedures based on sound mathematical evidence. Whenever the forester speaks of inventory involving volume estimates, of variations within a stand, of the effect of one treatment versus another, of the effects of silvicultural practices, he is employing statistical terms.

The presentation of data is one important application of statistical knowledge. Today, it is common practice to present data in an easily understood manner; in the newspapers or weekly news journals, we see the analysis of business trends, the division of the dollar into different parts to help pay for the cost of local, state and federal government expenditures or estimates of future population or industry. The information is presented in such a manner that the average person can readily grasp its meaning. Similarly, the forester must be able to reduce a mass of field data to a series of charts or graphs which give meaning to the data. The forester must not lose sight of the fact that the chart or graph is but a means to an end - to assist in making intelligent plans based on other statistical evidence.

In a course dealing with statistical techniques, formulas must be learned thoroughly and their application to practical problems understood. Formulas are nothing more than the result of a shorthand method of expressing various relationships; while they may seem complicated at first, diligent study and application of formulas to problem analysis will overcome the difficulty.

Research in forestry involves the planning of experiments, carrying out the provisions of the experiment in an efficient manner and analyzing the relationships which occur. Experiments may be relatively simple or very complex depending upon the nature of the problem. The design of the experiment is certainly important, as is the method of analyzing the results based on proven, acceptable statistical standards. In a text such as this, it is not possible to present all the possible experimental designs which might be used in forestry. Obviously, it would require a much more lengthy treatment. The object has been to present material which will be valuable to the profession as a whole, rather than to concentrate on a small, but important segment of it.

The scope of this book is sufficient to cover a full 3 semester credit hour course in statistical techniques, or the equivalent number of hours on the quarter system. The book is not intended to turn out expert statisticians; this it can not do. The purpose is to bring students and others to an appreciation of the value of statistical techniques in forestry.

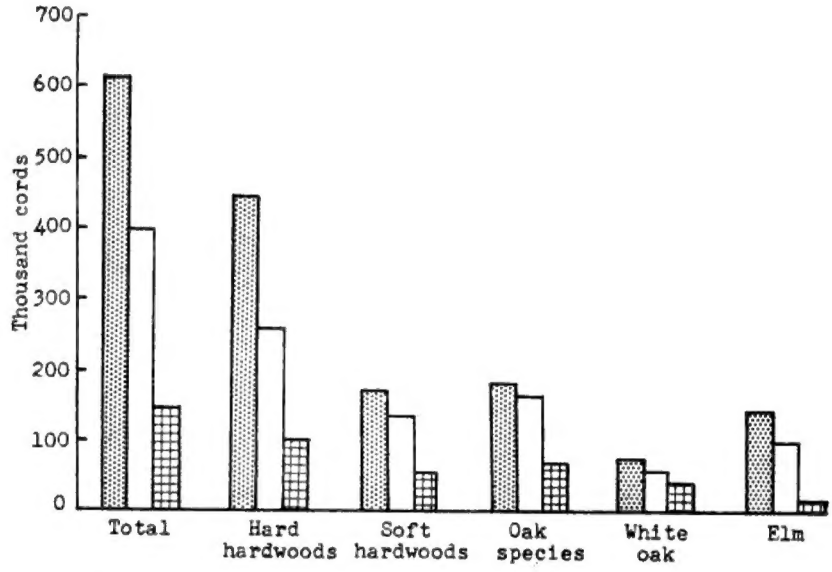
Chapter 1

GRAPHIC PRESENTATION OF DATA

The presentation of data in graphic form has become an essential part of a forester's work; it is his business to collect data, analyze them and to present the results in an easily understood manner. Data in tabular form are sometimes necessary in order to get detailed information, but for quick understanding a "word picture" is best.

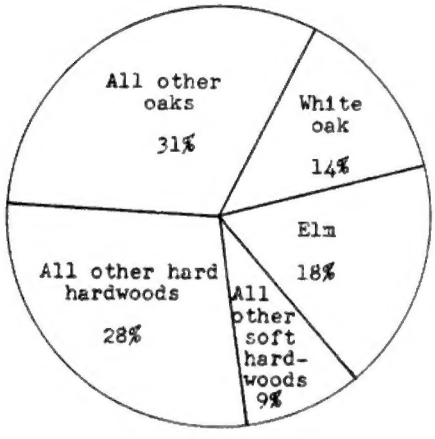
Graphic presentation may be accomplished in a number of ways; among the most frequently used are bar graphs, pie charts and line graphs. Each has its particular merits but emphasis will be on the construction and testing of line graphs.

Examples of bar charts and pie charts are shown in Figures 1.1 and 1.2.



(From Timber Resources of the Missouri Prairie Region, Bulletin B797, Univ. of Mo.)

Figure 1.1 - Example of a bar chart.



(From Timber Resources of the Missouri Prairie Region, Bulletin B797, Univ. of Mo.)

Figure 1.2 - Example of a pie chart.

Line Graphs

In its simplest terms, a line graph is a specialized visual aid to help us determine the relationship between two variables or factors; the relationship may assume any shape - a straight line, a simple curve or a complex one.

A variable is any factor which is allowed to change in value. Two types of variables are recognized: -

1. discrete
2. continuous

Discrete variable

When a variable can take on certain values only, it is known as a discrete variable. For instance, the number of students in the sophomore class represents a discrete variable because the variable can only assume integral numbers from 0 to some maximum; there can not be a fractional number of students, therefore the variable can assume certain values only, not all. The same argument can be used for other discrete variables - the number of pulp mills in a state, the number of seedlings in a nursery bed, the number of automobiles registered in a city, etc.

Continuous variable

On the other hand, a continuous variable can assume any value whatsoever within limits of minima and maxima which we might wish to impose. The only limitation is, in a sense, a physical one determining the accuracy of measurement. Take weight as an example; we could weigh ourselves on a machine having a scale divided into one pound intervals. On another type of scale, the weight might be more precise, being graduated into 1/10 lb intervals. Continuing this line of thought, an analytic balance weighs with an accuracy of 0.0001 gram. Weight, therefore, can assume any value. Other examples of a continuous variable are height, diameter, length, time and temperature.

Measurements give rise to continuous data while counting or enumeration gives rise to discrete data.

Line Graphs in Forestry

A great many relationships in forestry can be expressed in graphic form, either as straight lines or as curved relationships of one type or another. Forest mensuration deals not only with the measurement of data but also with the presentation of the data in as intelligent a manner as possible.

In a discussion of line graphs, early mention should be made of the distinction between an independent and a dependent variable. No rigid rules exist for the separation and, at times, the terms may be used interchangeably depending on the emphasis desired. However, the general rules are: -

1. a dependent variable is one which is to be estimated as a result of a measurement of another variable.

2. an independent variable is one which is to be used to provide the estimate.

One common problem in forest mensuration is to portray the relationship between total height in feet and diameter at breast height[1/]. The accepted procedure is to obtain data for each variable and to estimate total height in feet for a particular measured value of diameter in inches. Thus height becomes the dependent variable and diameter, the independent. By convention, the X-axis or abscissa is graduated in units of the independent variable, while the Y-axis or ordinate, is graduated in units of the dependent variable.

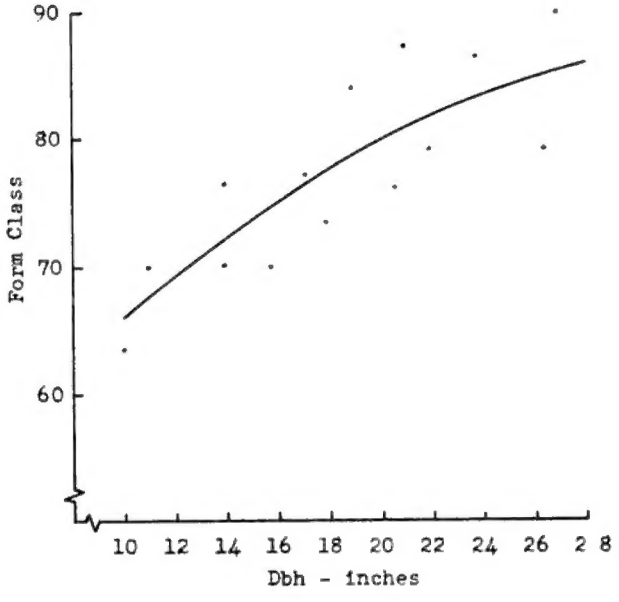


Figure 1.3 - Illustration of amputated axes.

Amputated Graphs

While it is good practice to utilize as much of the graph paper as possible, circumstances sometimes call for one or both axes of the graph to be amputated.

If the X-axis contained values from 10 inches to 30 inch dbh, this range is normally considered to be that of saw-timber trees. It is apparent that trees having a dbh smaller than 10 inches do not conform to the definition of saw-timber trees, so can be excluded from the data. The X-axis may be amputated below 10 inches dbh, leaving more room and greater opportunity to show the relationship between diameter and the second variable to better advantage.

[1/] Diameter at breast height is abbreviated to dbh; this abbreviation will be used throughout the text.

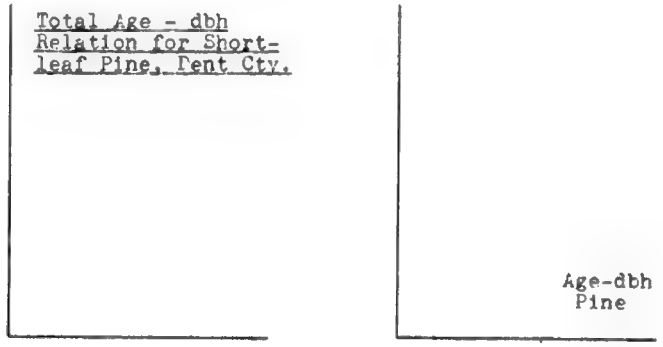
As an another example, form class is defined as the ratio between the diameter inside bark at 17.3 feet above ground to diameter outside bark at dbh. Form class is very useful in estimating changes in taper in trees above a minimum diameter, but has little application for smaller trees. The X- and Y-axes could both be amputated to eliminate wasteful and unnecessary work; amputation of both axes is shown in Figure 1.3.

The Complete Graph

Graphs are an integral part of forestry; so many relationships are shown in graphic form that it is almost second nature for a student or research worker to ask himself "Can this be shown graphically?" Too often, a graph is drawn without thinking of the person who will use it as a visual aid for a better understanding of the text to which the graph refers.

To achieve completeness in a graph, it should have: -

1. a title. The title is necessary for the identification of the relation between X and Y; the title identifies the problem being presented. Titles need not be long; they should be clear and to the point, however, and should be placed on the graph so that the title is almost the first thing to be noticed. For instance, compare these two titles side by side: -



a. Correct titling b. Incorrect titling

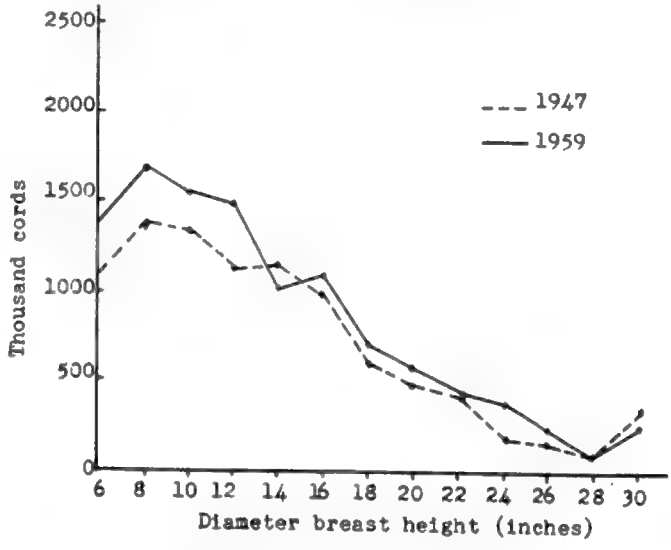
Figure 1.4 - Correct and incorrect methods of titling graphs.

The title in Figure 1.4a tells a story; its position attracts the eye and its wording is concise yet sufficient. Compare this with the title in Figure 1.4b.

In publications, the title is usually printed at the bottom of the graph as shown in Figure 1.5.

2. the two axes labelled. Each axis denotes classes or values of a particular variable, therefore each axis must be labelled so that there is no ambiguity as to its meaning. It is hardly necessary to point out that a graph is absolutely useless if the axes are not labelled at all.

The single word "height" as an axis designation could mean a number of things such as "total height in



(From Timber Resources of the Missouri Prairie Region, Bulletin B797, Univ. of Mo.)

Figure 1.5 - Distribution of the volume of growing stock by tree-diameter classes, 1947 and 1959.

feet", "merchantable height to a 4" top diameter", "number of 16' logs" and so on. Would it not be better to leave no room for doubt and to be precise in your wording?

3. a project or work reference. Because the project is so obvious to the person drawing the graph, it is often omitted. The reference is particularly important when many graphs are to be included in a report.

   Some projects are of a continuing nature in that data may be collected and analyzed each year and the results filed with the previous year's work.

   One problem which arises frequently is that a man may be moved to a different job, leaving unfinished work. You may be asked to continue the work and to extract as much information as possible from past records. Consider your frustration if a file contained notes and graphs, some of which had project references and some did not. Do you assume they all belong to the same project or not? The reference need only be brief, such as in Figure 1.6.

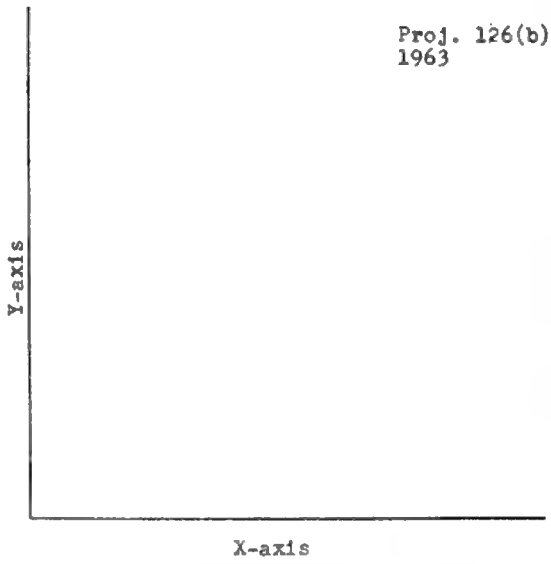


Figure 1.6 - Example of a project reference on a graph.

4. a source of information reference. A reader must be able to extract sufficient information from a graph that he need not thumb through pages of written material to find the answers to questions such as "Where did the data come from?" "How many measurements are involved?" "Is this the result of field work?" "Is this graph from some published source?" Again, a brief and concise statement will suffice.

5. initial and date reference. Lastly, we come to the author of the graph. Who is he? When did he do the work? A person drawing a graph should have pride in his work and should sign his work with his initials and date. Some foresters use this system on all graphs they draw, whether they are rough drafts or the completed product.

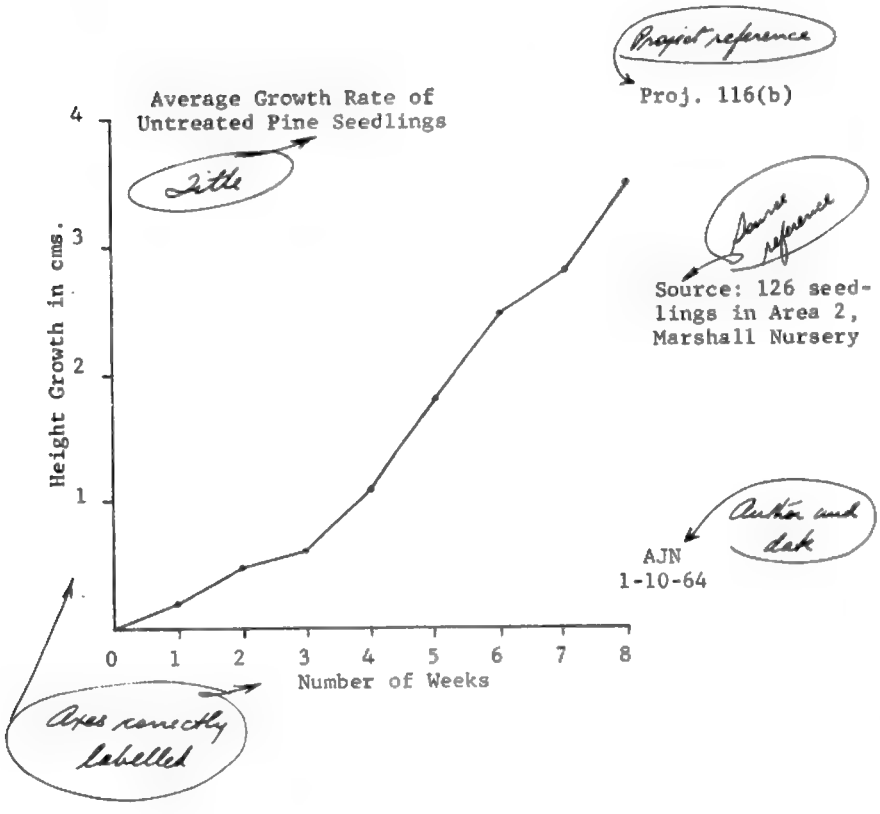


Figure 1.7 - The complete graph.

An example of the five points discussed is shown in Figure 1.7.

The approximate locations of the title, work reference, source and author references are indicated in Figure 1.7; these are not rigidly defined but experience has shown preference to this arrangement over others.

These hints on the drawing and presentation of line graphs are made so that your work may be done in a manner acceptable to the forestry profession.

Freehand Method of Fitting a Curve

The process of drawing a curve or straight line which will best fit points plotted on two axes is known as "estimating by means of two variables". There are two acceptable methods; the first is a "freehand" method which does not resort to mathematical proof, whereas the second uses formulas and gives an exact mathematical expression of the relationship between the two variables. Only the freehand method will be discussed in this chapter; the presentation of the mathematical method is being postponed until Chapter 8.

The term "freehand" is an unfortunate one because it implies that the final curve is completely freehand; this is seldom true. Perhaps the term "non-mathematical" would be more appropriate.

Starting with a list of paired values - total height in feet associated with particular dbh values for instance - points are plotted on graph paper to correspond with the paired values. If the points lie in a fairly tight band, you will have no difficulty in determining the trend of the curve. If the points are scattered, the trend might not be apparent and it might be advisable to resort to a mathematical solution. Let us assume, however, that points are plotted and the trend is evident. The procedure for completing the curve is as follows: -

1. draw a trial curve by hand so that there is an approximate balance between the number of points above and below your trial line.

   (A useful tip on drawing reasonably smooth freehand curves is to place the sheet of graph paper so that the curve can be drawn away from you. By pivoting your arm on your elbow, you can use short, light strokes to position the first trial curve. For curves which start in the upper left corner and end in the lower right corner, turn the paper around so that the upper edge of the graph paper is toward you. Start drawing the curve where the X-values are high. This will result in a curve known as a "reverse-J" and is typical of the relationship between number of trees per acre and dbh. Do NOT use a soft pencil; a 3H or 4H one is best for curve drawing.)

2. determine the total of the positive deviations. A positive deviation is one in which the Y-value of the point is greater than the curve value for the same value of X. The deviations are always taken in the Y- direction. Figure 1.8 illustrates positive and negative deviations.

3. determine the total of the negative deviations in a similar manner.

4. subtract the smaller from the larger.

5. if the difference obtained in (4) is within acceptable limits (established ahead of time), complete the curve by drawing it in final form with a French curve or similar drawing device.

6. if the difference obtained in (4) is not within acceptable limits, the trial curve must be moved. It should be moved down if the negative deviations total more than the positives and vice versa.

7. repeat the process until balance is achieved and complete the curve with a French curve.

If the data consist of a large number of observations covering a wide spread of X values, it might be desirable to group the data by classes. A point on the graph will now represent the average X and average Y for a particular class of X; since each point now represents an average, it

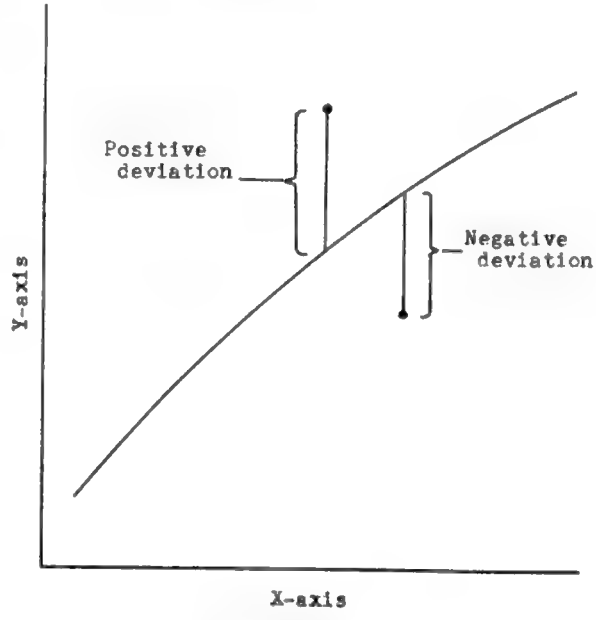


Figure 1.8 - A positive and negative deviation shown graphically.

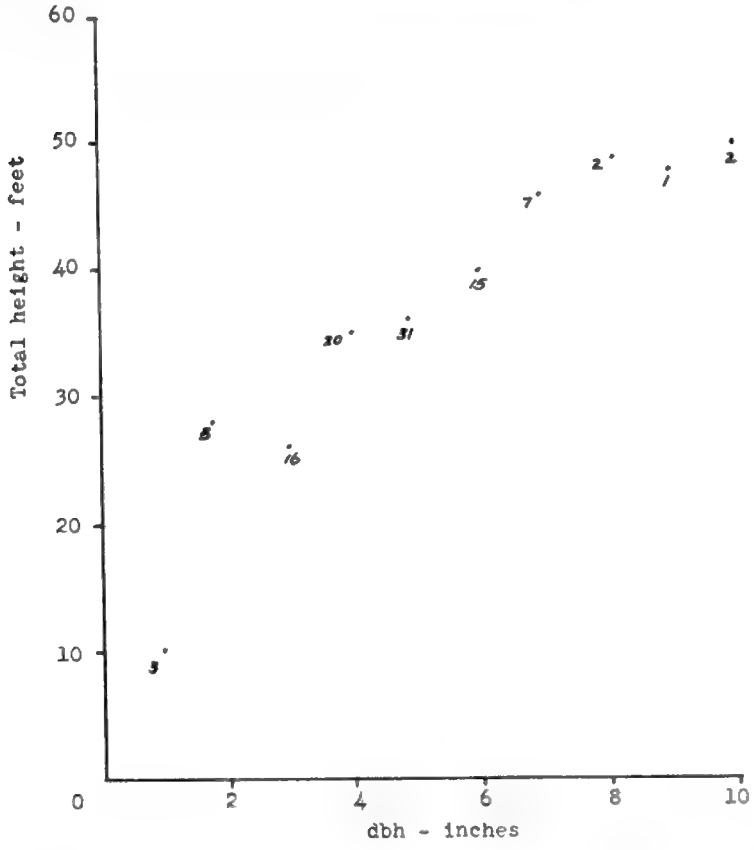


Figure 1.9 - Plotted points with frequency of each shown.

must have come from a number of observations. The frequency MUST be entered beside each point so that when approximate balance is being sought, the curve will pass close to the points of high frequency. Balance is achieved when the sum of the positive deviations multiplied by their individual frequencies equals the sum of the corresponding negative frequencies.

All curves in forestry are SMOOTH and do not have illogical bends in them. A great number of curves in the normal type of forestry problem can be freehand curves and it is essential that you master the technique of producing acceptable work. It might take two or more trials to get a curve balanced; do not be discouraged since it takes a lot of practice to position a curve correctly and even more practice to draw it smoothly.

Example of drawing a freehand curve

The data given below in Table 1.1 are to be used in the construction of a height-diameter curve. A number of trees in each dbh class was measured and the heights and diameters shown are averages within each class. Let us follow the procedure for making the curve step by step.

Table 1.1 - Data for height and diameter measurements of 102 white oak trees measured in southern Illinois, 1960. Project 12.

dbh inches	Total height feet	Number of trees
1.0	10	3
1.9	29	5
3.0	26	16
4.0	35	20
4.9	36	31
6.0	40	15
7.0	46	7
8.1	49	2
9.0	48	1
10.0	50	2
		102

The procedure in drawing the curve is: -

1. plot the points as shown in Table 1.1. Use a scale on each axis so that as much of the graph paper as possible will be used. Label the axes correctly and indicate the frequency of each point. See Figure 1.9 at the top of this page.

2. Sketch a trial curve by eye, starting at 4.5 feet on the height axis, since the diameter measurements are recorded at dbh, not ground level or any other point on the stem. Make sure that your first trial curve passes close to the points of highest frequency. When you have it positioned to the best of your ability, record the positive and negative deviations. These points are shown in Figure 1.10 and Table 1.2.

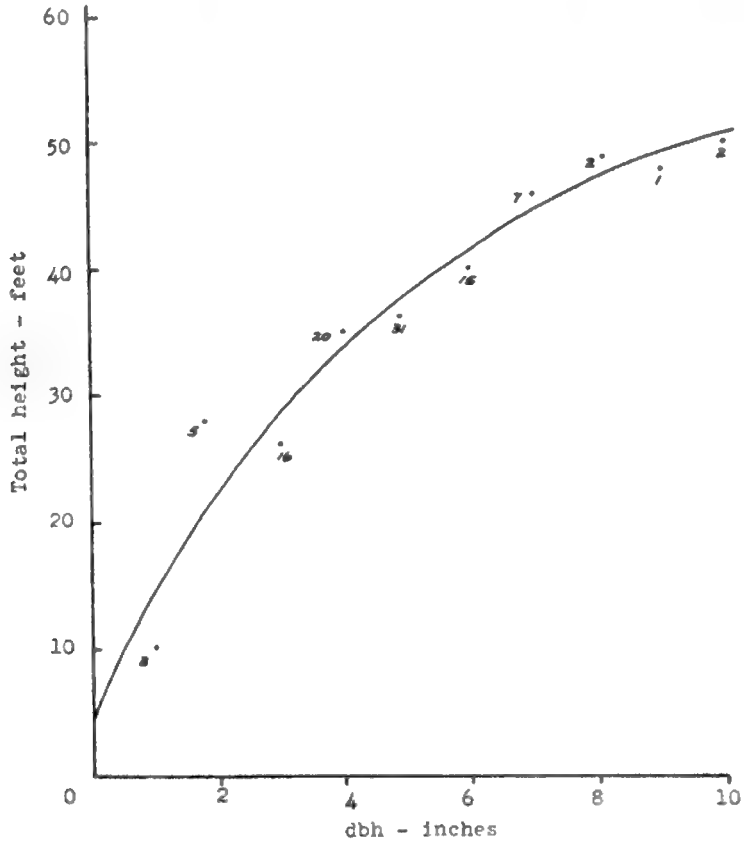


Figure 1.10 - First trial curve of height on diameter.

Table 1.2 - Record of positive and negative deviations for the first trial curve.

dbh	Total height	Number of trees	Deviations		(Frequency)(Deviations)	
inches	feet		+	-	+	-
1.0	10	3		2.0		6.00
1.9	29	5	3.5		17.5	
3.0	26	16		1.5		24.00
4.0	35	20	1.0		20.0	
4.9	36	31		0.75		23.25
6.0	40	15		0.75		11.25
7.0	46	7	0.5		3.5	
8.1	49	2	1.0		2.0	
9.0	48	1		0.75		0.75
10.0	50	2		0.50		1.00
					43.0	66.25

3. In order to be balanced, the positive and negative deviations should balance exactly; however, it is usual to allow a slight margin and to consider a curve of this type balanced if the difference between the positive and negative deviations is not greater than 2.0. Table 1.2 shows that the difference is far above the allowable limit, so the curve must be adjusted. Since the negative deviations exceed the positives, the curve must be lowered. Again, this new position of the curve can be sketched in and a new computation of the deviations made. This process is repeated until the curve is balanced within the allowable error. Figure 1.11 shows the position of the adjusted curve, with Table 1.3 showing the deviation computations.

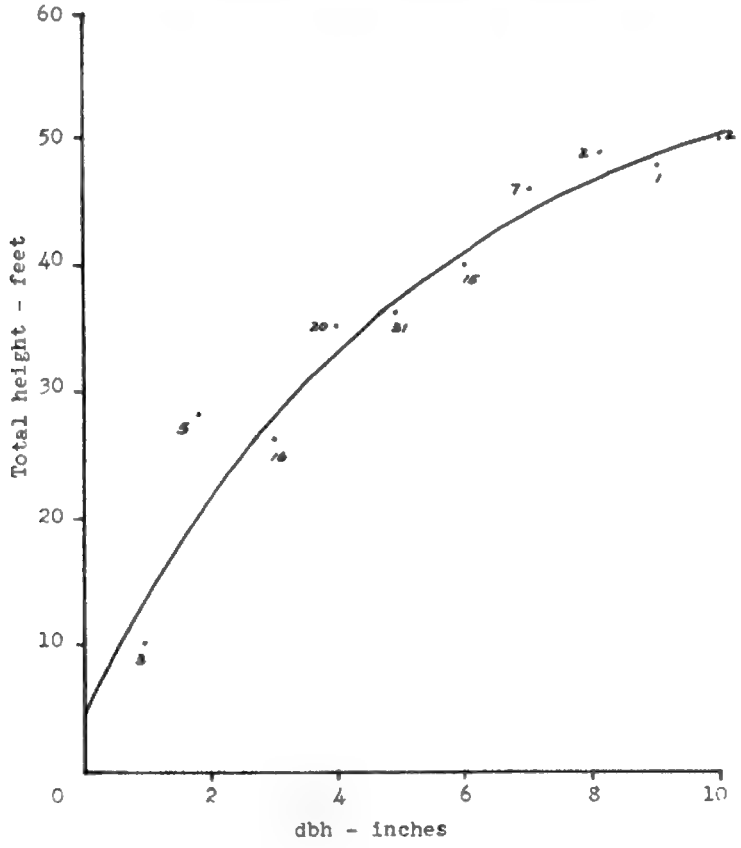


Figure 1.11 - Second trial curve of height on diameter.

Table 1.3 - Record of positive and negative deviations for the second trial curve.

dbh	Total height	Number of trees	Deviations		(Frequency) (Deviations)	
inches	feet		+	-	+	-
1.0	10	3		2.0		6.0
1.9	29	5	4.0		20.0	
3.0	26	16		1.0		16.0
4.0	35	20	0.75		15.0	
4.9	36	31		0.50		15.5
6.0	40	15		0.50		7.5
7.0	46	7	1.0		7.0	
8.1	49	2	1.0		2.0	
9.0	48	1		0.50		0.50
10.0	50	2	0.25		0.5	
					44.5	45.5

The second trial curve was successful in achieving balance because the difference between the positive and negative deviations was 1.0.

4. Now that the position of the curve has been established, it can be drawn in final form with a French curve and the graph completed. This is shown in Figure 1.12. The work will be finished when you make a table showing the estimated height read from the curve, for each integral inch of dbh; Table 1.4 gives these data.

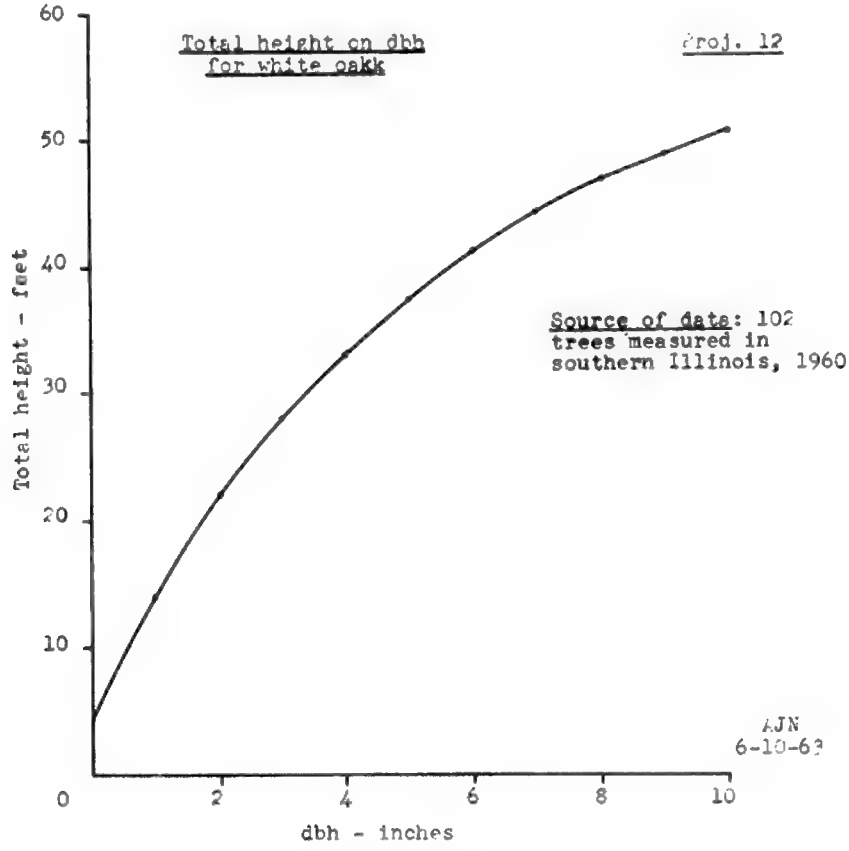


Figure 1.12 - Completed graph of height on dbh for white oak measured in southern Illinois, 1960.

Table 1.4 - Total height in feet for integral inch classes of dbh for white oak in southern Illinois.

dbh	Total height
inches	feet
1.0	14
2.0	22
3.0	28
4.0	33
5.0	37
6.0	41
7.0	44
8.0	47
9.0	49
10.0	51

The data in Table 1.4 can now be used for any purpose in which total height in feet for inch classes of dbh is required; the most common application of height-diameter data is for the computation of volumes of individual trees.

Exercise 2 in the Appendix is designed to give you practice in the drawing and balancing of freehand curves.

Chapter 2

CLASS INTERVALS AND SOME FREQUENCY DISTRIBUTIONS

Class Intervals

In Forest Mensuration, Forest Management or in other areas of forestry, we want to establish the frequency of occurrence of a measured variable. Look at the following list: -

dbh inches	Frequency
2	3
3	8
4	24
5	36
6	21
7	10
8	4

It establishes the fact that there are three trees with a dbh of two inches, eight at three inches, etc. But are the trees in the first category all exactly two inches at dbh? Or does the designation "two inches" represent the mid-point of a class? More than likely it does; very seldom do all trees in a stand or group measure exactly the same. Obviously then, we are dealing with the mid-points of class intervals rather than exact measurements.

The raw data in the foregoing list could have been shown as 1.7, 2.4 and 2.1 inches for the first class, 2.6, 2.9, 3.5, 3.2, 2.8, 3.2, 3.1, and 2.6 inches for the next class and so on.

The mid-points of class intervals are very convenient ways of listing data. But what are the limits of each class? It seems reasonable to set the limits of a class as being half-way between class mid-points. For instance: -

Mid-point inches	Class Limits Lower inches	Upper inches
2	1.5	2.5
3	2.5	3.5
4	3.5	4.5
5	4.5	5.5

etc.

Confusion arises when a measurement falls at the exact lower or upper class limit. A dbh of 3.5 inches can logically be placed into either the 3-inch class (upper limit) or into the 4-inch class (lower limit). So far, we do not have rules for separating the two, except to say that a class limit should not coincide with an actual measurement. In forestry measurements, this objective is impossible to achieve. So we arbitrarily decide that for purposes of convenience, the class limits will not be at the exact mid-points. For the 1-inch class that was illustrated, the limits are: -

Mid-point inches	Class Limits Lower inches	Upper inches
2	1.6	2.5
3	2.6	3.5
4	3.6	4.5
5	4.6	5.5

etc.

For 2-inch classes, listed as 2-, 4-, 6-, 8- inches and so on, the limits would be: -

Mid-point inches	Class Limits Lower inches	Upper inches
2	1.1	3.0
4	3.1	5.0
6	5.1	7.0
8	7.1	9.0
	etc.	

The same reasoning can be applied to measurements and classes of height, age, length or any other variable we may wish to measure. The fore-going discussion of class intervals and class limits is restricted to the disposition of measurements which are to be separated into classes. It does not refer to classes of a variable which can be treated as continuous for the purposes of defining statistics such as the median, mode and the mean. (See Chapter 3)

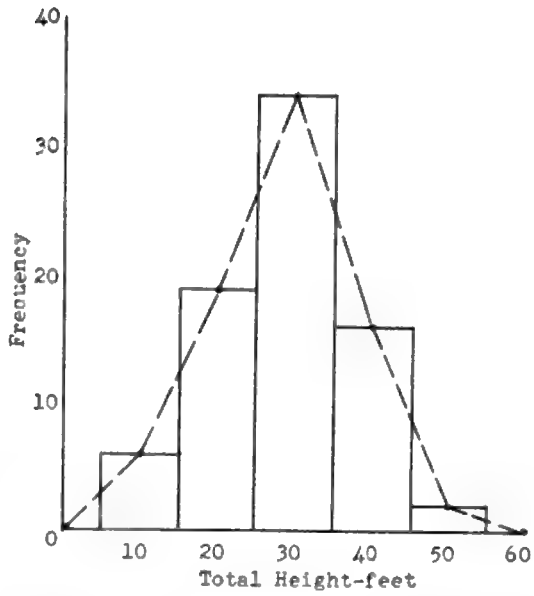


Figure 2.1 - Histogram and frequency polygon.

Frequency Distributions

When frequency is plotted on a variable, we can show the results as: -

1. a histogram
2. a frequency polygon
3. a cumulative frequency polygon or ogive

Consider the following example: -

Total height feet	Number of trees
10	6
20	19
30	34
40	16
50	2
Total	77

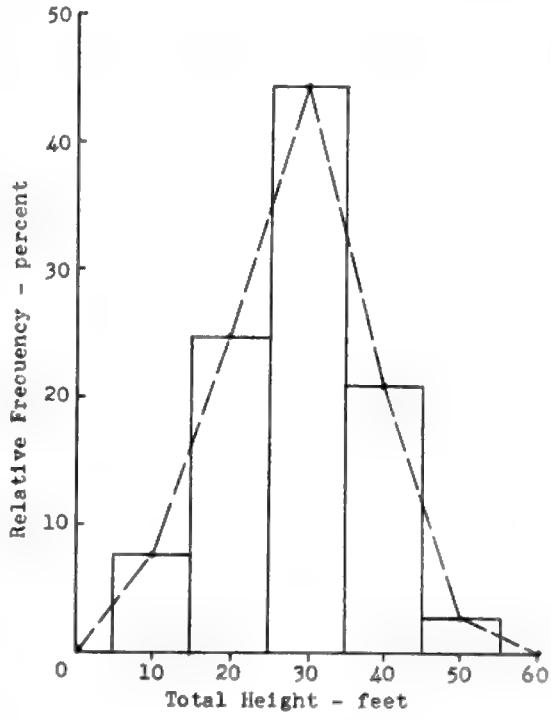


Figure 2.2 - Relative Histogram and Relative Frequency Polygon.

A histogram depicts the frequency of each class as a rectangle having an altitude proportional to the class frequency. If the mid-points of the classes were to be connected by a continuous line, we would have a frequency polygon. These two are illustrated in Figure 2.1.

If the frequencies were computed as percentages of the total frequency, the graph would be exactly the same shape but would be known as either a relative histogram or a relative frequency polygon. Using the same data as before, the relative frequencies (percentages) would be as follows. -

Total height feet	Number of trees	Relative Frequency percent
10	6	7.8
20	19	24.7
30	34	44.2
40	16	20.8
50	2	2.5
	Total 77	100.0

Cumulative Frequency and Cumulative Relative Frequency Polygons

An interesting and important function of frequency distributions occurs when the frequencies are cumulated class by class, starting at the smallest class of X. As we progress further in the study of statistics, we will see the importance of cumulative frequency polygons, but for the present, we are only interested in learning how the trends are established. The titles "Cumulative Frequency Polygon" and "Cumulative Relative Frequency Polygon" are rather long and the term "ogive" has been given to the first and "relative ogive" to the second.

The original data are shown below, with the frequencies and relative frequencies cumulated.

Total Height feet	Number of Trees	Cumulated Frequencies	Relative Cumulated Frequencies percent
10	6	6	7.8
20	19	25	32.5
30	34	59	76.7
40	16	75	97.5
50	2	77	100.0
Total	77		

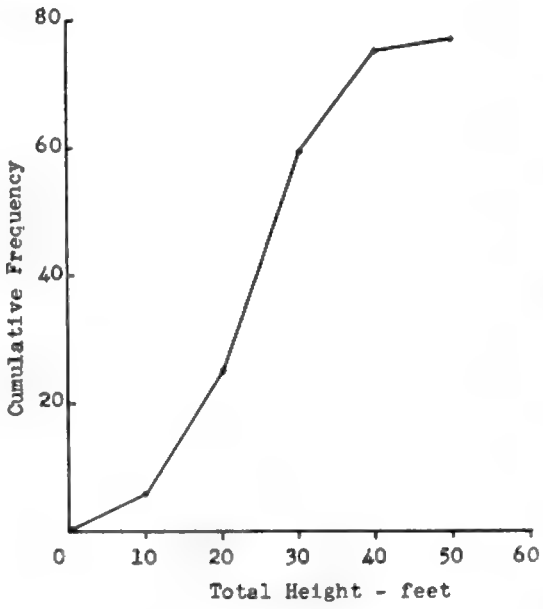


Figure 2.3 - Cumulative frequency polygon (ogive) of data given on page 11.

The ogive and relative ogive are shown in Figures 2.3 and 2.4 respectively.

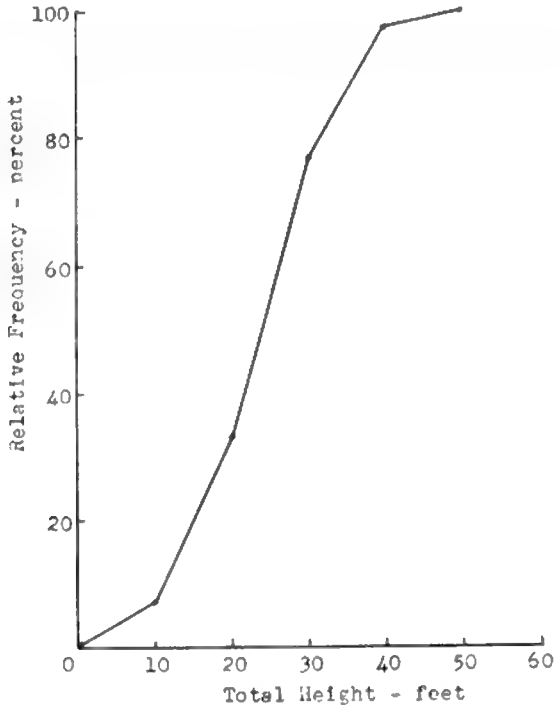


Figure 2.4 - Relative cumulative frequency polygon (relative ogive) of data listed on page 11.

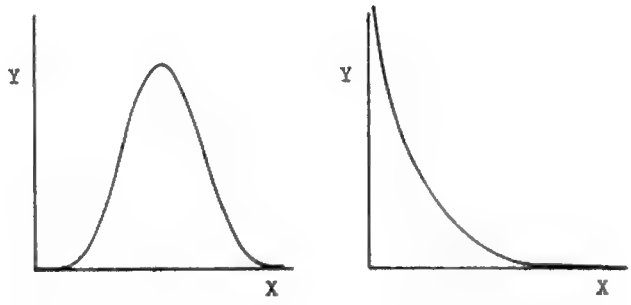
Frequency Curves

Instead of having classes of X which are quite large, we could think of classes which get progressively smaller and smaller. The frequency polygon would then have very small segments and would appear smoothed out. Drawing a smooth curve from a frequency polygon has the effect of reducing the size of the classes and treating the X variable as if it were continuous.

In forestry, we are concerned with two or three types of frequency curves. For the present, we are going to illustrate them without detailed explanation of either the mathematics involved or the application of the curves. The three types of curves are: -

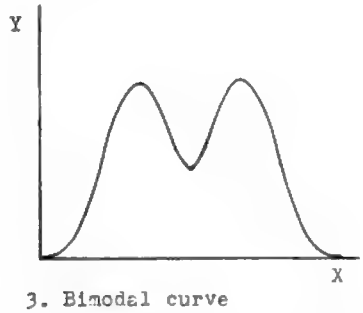
1. bell-shaped or normal curve
2. reverse-J curve
3. bimodal or double-humped curve

and are illustrated in Figure 2.5.



1. Bell-shaped or normal curve

2. Reverse-J curve



3. Bimodal curve

Figure 2.5 - Three types of curves having application in forestry.

Chapter 3

MEASURES OF CENTRAL TENDENCY

The previous two chapters dealt with the use of graphs and some general frequency distributions; these were necessary to lead us into a more specific discussion involving statistical notation and measures of central tendency.

Nearly all functions or relationships in the field of statistics are symbolized. Unfortunately, not all texts use the same symbols to denote the same thing. However, there has been a conscious effort in recent years toward standardization of symbols. In this treatment of statistics, symbols have been kept as simple as possible, but staying within the realm of convention.

It has become almost standard practice to denote variables such as age, height, weight or length as X, if only one is being considered. When two variables are being used, we call them X and Y; when three, X, Y and Z. In the case of independent and dependent variables, the independent variable is called X while the dependent is called Y, to conform to the axes of graphs. A specialized notation is used in more advanced work where there might be five or more independent variables; in this case, we use X_1, X_2, X_3 -----X_n. The letters at the beginning of the alphabet are usually reserved for constants or coefficients such as a, b, c, etc.

The mean as a measure of central tendency

If there were a list of height measurements of male University students, each height would be a measurement of the variable X; the sum of heights would be ΣX[1/], and the average or mean value of X would be written as $\bar{X}$, (X-bar).

As you are well aware, the mean of a number of items is the sum of them all divided by the number of items. In symbolized language, this becomes: -

$$\bar{X} = \frac{\Sigma X}{N} \quad \text{or} \quad \frac{\Sigma f x}{\Sigma f} \tag{3.1}$$

where $\bar{X}$ = mean value of X

ΣX = sum of the X values

N = number of items when each individual is listed or counted separately

Σf = total number of items when each has a frequency of 1

Thus, the heights of 20 male University students, measured to the nearest one inch, might be 68, 72, 66, 69, 63, 74, 70, 69, 67, 69, 70, 64, 74, 67, 67, 69, 74, 71, 66 and 67. The sum of all heights (ΣX) is 1376 inches and the mean value of X is: -

$$\bar{X} = \frac{1376}{20} = 68.8 \text{ inches.}$$

We have established our first measure of central tendency, the mean.

Some of the heights are repeated, while others occur only once in the above list; we could take each height, starting with the lowest value, and record the number of times each height occurs. For the sake of continuity, X values of 65 and 73 are included even though they do not occur in the original list. They have a frequency of 0.

1/ Σ is Greek capital sigma and is read "the sum of . . . "

Multiplying each X by its class frequency gives us an fX, the sum of which is ΣfX. Therefore, when data are grouped into classes having different frequencies, the formula for the mean becomes: -

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} \qquad (3.2)$$

X	f	fX
63	1	63
64	1	64
65	0	0
66	2	132
67	4	268
68	1	68
69	4	276
70	2	140
71	1	71
72	1	72
73	0	0
74	3	222
	Σf = 20	ΣfX = 1376

where $\bar{X}$ = mean value of X

ΣfX = sum of each X class multiplied by its particular frequency

Σf = sum of the class frequencies.

The value of the mean is $\frac{1376}{20}$ inches = 68.8 inches as before and the task of computing the mean has been reduced considerably.

If there is a large number of classes or the frequency in each class is high, the above procedure can become quite time-consuming. There is a short-cut method available; in this, we -

-assume a mean

Instead of calculating a mean directly, the data may be sorted into classes of X as before. Now, rather than perform laborious computations, we can choose any one of the X classes and call it the assumed mean. Indeed, it may not be one of the classes actually present, but can be any value whatsoever. The chances are remote that our assumed mean will be equal in value to the true mean. This does not matter, because a correction factor will take care of the difference. The formula for obtaining the true mean from an assumed mean is: -

$$\bar{X} = \bar{X}_A + \frac{\Sigma fx'}{\Sigma f} \qquad (3.3)$$

where $\bar{X}$ = true mean

$\bar{X}_A$ = assumed mean

Σfx' = the algebraic sum of the deviations of each class value from the assumed mean multiplied by the class frequency

Σf = sum of the class frequencies.

A word of explanation is necessary at this point concerning the symbol x'. A deviation (or difference) between any X value and the true mean is shown as $X - \bar{X}$ and is given the symbol

x, always a lower case letter. If, however, a deviation is desired between an X value and an assumed mean, (X_A), the deviation is shown as x' to distinguish it from x.

Returning to the height problem, let us assume a mean of 70 inches and calculate the true mean by Formula 3.3. The steps are: -

1. list the X values with their associated frequencies (Columns 1 and 2)
2. compute x' ($X - \bar{X}_A$) for each class and enter the result opposite each class (Column 3)
3. multiply each x' by the class frequency to obtain fx'. Enter this opposite each class (Column 4)
4. add the negative values of fx'; add the positive values of fx' and obtain the difference to get $\Sigma fx'$ which is the algebraic sum
5. using the assumed mean of 70 inches and the values of $\Sigma fx'$ and Σf, determine the value of the true mean.

(1) X	(2) f	(3) x'	(4) fx'	
63	1	- 7	- 7	
64	1	- 6	- 6	
65	0	- 5	0	
66	2	- 4	- 8	
67	4	- 3	-12	
68	1	- 2	- 2	
69	4	- 1	- 4	-39
70	2	0	0	
71	1	+ 1	+ 1	
72	1	+ 2	+ 2	
73	0	+ 3	0	
74	3	+ 4	+12	+15
	$\Sigma f = 20$		$\Sigma fx' =$	-24

The true mean is therefore,

$$\bar{X} = 70 - \frac{24}{20}$$

$$= 70 - 1.2$$

$$= 68.8 \text{ inches}$$

Our assumed mean of 70 inches was too high; we knew this because the true mean of this number of heights was 68.8 inches. But we need not have known the value of the true mean at all. The correction factor $\frac{\Sigma fx'}{\Sigma f}$ takes our wrong assumption into account. If the assumed mean had been lower than the true mean, the sum of the positive fx' values would have been greater than the sum of the negatives, thereby raising the value of the assumed mean.

Assuming a mean has its advantages in the calculation of $\bar{X}$ but later sections will show that even greater advantages will be demonstrated in the calculation of other statistics.

The median, a second measure of central tendency

In addition to the mean, we have another measure of central tendency; it is the median, which occupies the middle value of an array of values when they are arranged in order of

magnitude. It is the value of X which divides the frequency distribution into two equal parts. The median is affected by the number of items only, not by their sizes.

The formula for determining the median value is: -

$$X_{me} = L_{me} + \frac{i}{f} \quad (C) \tag{3.4}$$

where X_{me} = median

L_{me} = lower limit of the class in which the median is located, assuming the variable to be a continuous variable

i = the difference between the mid-frequency (N/2) and the cumulated frequency of all classes lower in value than the median class

f = frequency in the median class

C = class interval.

We will determine the median value of the following distribution: -

X	f	Cumulative Frequencies
0	4	4
1	10	14
2	18	32
3	35	67
4	22	89
5	17	106
6	6	112
	Σf = 112	

Half the total frequency is 112/2 or the 56th item; this must occur in the class where X = 3 since there is a total of 32 items below the median class. The median item occurs at the point in the median class where X is the 56th item in the whole array. According to the formula,

$$X_{me} = 2.5 + \frac{24}{35} \quad (1)$$

$$= 2.5 + 0.68$$

$$= 3.18$$

Quite often in economic statistics, we hear of the median salary being $6700; it means that as many workers in the category under study are earning salaries below $6700 a year as are earning more.

Another formula which is also used to compute the median, and which gives exactly the same results is: -

$$X_{me} = L_{me} + \left(\frac{N/2 - \Sigma f_{cum}}{f_{me}} \right) \cdot C \tag{3.5}$$

where X_{me} = median value

L_{me} = lower limit of the median class

N = total frequency

Σf_{cum} = the sum of all the frequencies in classes lower in value than the median class

f_{me} = frequency in the median class

C = class interval.

We will meet the median again later in our discussion of the normal frequency curve.

The mode, a third measure of central tendency

A third measure of central tendency is the mode which is defined as the most frequent item or the value of X which occurs most often in an array. If frequency were plotted on X, the mode would be value of X at which the peak of the curve occurs. Sometimes, there is no modal value and at others, there might be two or more modes. The first of these would occur in a curve such as a "reverse-J" and the latter in either a bimodal or a multi-modal curve. These are particular cases.

Generally speaking, there is a value of X which occurs most often in an array, and the formula for determining it is: -

$$X_{mo} = L_{mo} + \left(\frac{f_H}{f_H + f_L} \right) \cdot C \qquad (3.6)$$

where X_{mo} = modal value

L_{mo} = lower limit of the class in which the mode occurs, assuming the variable to be continuous

f_H = frequency of the class immediately higher in value than the modal class

f_L = frequency of the class immediately lower in value than the modal class.

Let us determine the value of the mode in the following frequency distribution: -

X	f
0	0
1	3
2	8
3	15
4	25
5	50
6	40
7	30
8	20
9	5

The modal class is where X = 5 since it contains more items than any other class. X_{mo} will be some value of X toward X = 6 because of the frequencies in the two classes on either side of the modal class. The frequency of the class immediately higher in value than the modal class is 40, while that immediately lower is 25, therefore the mode will be pulled toward the higher frequency. Computing the mode from the formula gives us: -

$$X_{mo} = 4.5 + \left(\frac{40}{40 + 25}\right) \cdot 1$$

$$= 4.5 + \left(\frac{40}{65}\right) \cdot 1$$

$$= 4.5 + 0.615$$

$$= 5.115$$

Quantiles

The word "quantile" is a general term meaning a particular but unidentified number in a distribution which has been ranked in order.

To be more specific, we may think of the terms quartiles, deciles and percentiles as dividing the distribution into four, ten and a hundred equal parts respectively. The quartiles, for instance, are numbers representing the $\frac{Nth}{4}$, $\frac{2Nth}{4}$, $\frac{3Nth}{4}$ items in the distribution; likewise, the deciles are the $\frac{Nth}{10}$, $\frac{2Nth}{10}$, $\frac{3Nth}{10}$ ..... $\frac{9Nth}{10}$ items. The quartiles are symbolized as Q1, Q2 and Q3, the deciles by D1, D2, D3.....D9 and the percentiles by P1, P2, P3.....P99. Thus the numerical value of Q2, D5 and P50 are equal and also represent the median value of the distribution.

Let us take the following distribution and calculate the various quantiles.

Hours worked per week	Number of workers	Cumulative frequency
0 - 9.99	2	2
10 - 19.99	6	8
20 - 29.99	24	32
30 - 39.99	41	73
40 - 49.99	20	93
50 - 59.99	6	99
60 - 69.99	1	100

Q1: $\frac{N}{4} = \frac{100}{4}$ = 25th item starting with the smallest

Q1 will occur in the class 20 - 29.99 since the cumulative frequency up to that class is 8; we need (25 - 8) or the 17th item in the class. By straight proportion

$$Q1 = 20 + \frac{17}{24}(10.00) = 20.00 + 7.08$$

$$= 27.08$$

Q2: $\frac{2N}{4} = \frac{200}{4} =$ 50th item

Q2 will occur in the class 30.00 - 39.99 because the cumulative frequency up to that class is 32. So we need (50 - 32) or the 18th item in the class.

$$Q2 = 30.00 + \frac{18}{41} (10.00)$$

$$= 30.00 + 4.39$$

$$= 34.39$$

Q3: $\frac{3N}{4} = \frac{300}{4} =$ 75th item. By the same process

$$Q3 = 40.00 + \frac{2}{20} (10.00)$$

$$= 40.00 + 1.00$$

$$= 41.00$$

The interpretation of these quartiles is that (a) 25% of the workers worked less than 27.08 hours per week (b) 50% worked less than 34.39 hours per week and (c) 75% worked less than 41.00 hours per week.

The deciles can be calculated in a similar manner, as can be the percentiles.

Chapter 4

POPULATIONS, SAMPLES AND MORE ON FREQUENCY DISTRIBUTIONS

Population

In statistical language, a population is made up of all the items occurring within a given definition. A population can be theoretically infinite in size, with any value of the measured variable being possible, or it may be finite in size. The application of statistical theory assumes that we are dealing with infinite populations, but practically speaking, these seldom occur. For instance, if we defined a particular population as including all male University students within a state, the number of students is large but is of a definite - or finite- quantity. Again, we might take all trees of a certain species within a forest district; the number is large, but not infinite.

A population must be precisely defined because, as we shall see shortly, we will use samples taken from a population to estimate various properties of the population. The population must be defined in such a manner that our samples are taken from the specific population and no other. If, by error or by failure to define a population correctly, samples were taken from a mixture of populations, the results would not be very meaningful in terms of the population we had originally intended.

Examples of correctly-defined populations are: -

1. fully-stocked white oak stands in Missouri.
2. monthly production of parts by Machine X at the ABC Metal Co.
3. all shortleaf pine sawtimber trees on site 80 on a Clarksville soil series.
4. 0 - 2 loblolly pine seedlings in a nursery raised from Arkansas seed.
5. farm enterprises within a state which have gross incomes between $10,000 and $15,000 annually.

Seldom, if ever, is every member of the population measured for a particular characteristic. If this were true, we would have no further need to study statistics because we would know everything possible about the population. Since we do not measure every unit or member, we rely on the results of samples so we can deduce certain things; hence the term "deductive" statistics which means that we go from the particular (the sample) to the general (the population).

Samples

Samples are so important in statistical techniques that they deserve special mention. A sample is a representative group of items drawn from a population (1) to furnish information on the variability of the items making up the sample and (2) to permit estimates of population parameters such as the mean (μ), the standard deviation (σ) and others which will be discussed in Chapter 5. The same terms applied to a sample are called "statistics". Our main interest lies in the population rather than in the sample, which is a means to an end.

In forest inventory work, samples are taken in the field and consist of measurements on small areas scattered throughout the forested area defined by the population. From the sample, we can determine, within quite precise limits, complete inventories of volume. On large inventories, such as on a regional basis, the total area measured by the sample may be as little as 0.1% whereas sampling on a 160-acre forest may cover as much as 10% of the area.

Since sampling is so important, it must be done correctly. Each sample must satisfy certain requirements which are: -

1. the sampling units - the individual items which are measured, whether it is height, volume, number of trees, size of parts etc. An unbiased sample implies that the estimates of population parameters will be unbiased. Personal selection should not enter into the sampling process; human nature being what it is, tends to make us select the better individuals at the expense of those which are "below average", resulting in an estimate which is too large in value. Methods of selecting samples are discussed in Chapter 7.

2. the number of sampling units chosen must be sufficient to ensure that the population is represented. No rigid requirements can be laid down to ensure adequate representation. A population which consists of items which vary very little between them can be sampled with a relatively small number of sampling units. One exhibiting wide variation must be sampled more thoroughly in order to detect this wide variation. A special phase of statistical study has been devoted to the effects of small samples where the number of sampling units is less than 30.

The normal curve

One of the most important distributions in statistical work is the normal or Gaussian distribution; it is a mathematical relation expressed by the formula: -

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-\mu)^2}{2\sigma^2}} \qquad (4.1)$$

where μ = mean value

π = 3.14159

e = 2.71828

σ = standard deviation

Y = value of the ordinate

The shape of the curve is illustrated in Figure 4.1.

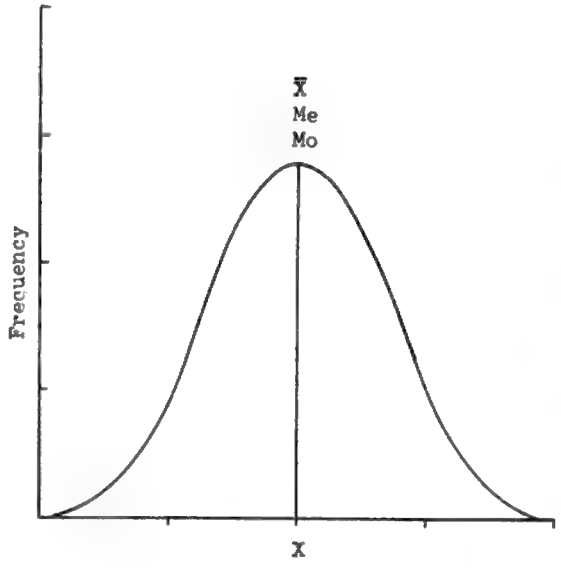


Figure 4.1 - Normal curve with $\bar{X}$, Me and Mo identical in value.

A normal curve has certain characteristics which are unique: -

1. the mean, median and mode are identical in value.

2. the curve is perfectly symmetrical on both sides of the mean. In other words, the frequency of occurrence of X values is the same for similar deviations from the mean. If the mean value were 8.6, the frequency for an X = 5.0 ($X - \bar{X} = -3.6$) is the same as for an X = 12.2 ($X - \bar{X} = 3.6$).

3. the X-axis theoretically goes out to $-\infty$ and $+\infty$

The X-axis for a normal curve is assumed to be continuous. If the X-axis denotes a discrete variable, a curve very similar in shape to the normal curve will result from plotting frequencies by expanding the binomial expression $(p + q)^n$ where p is the probability of success and q is the probability of failure in any single trial. As n increases, the curve resulting

from the binomial expansion very closely approximates a normal curve and is sometimes used to illustrate the method by which a normal curve is generated.

The distribution derived from the binomial expansion is often called a Bernoulli distribution after James Bernoulli who discovered it in the latter part of the 17th century.

Two examples of normal frequency distributions are: -

1. diameter distribution in an even-aged stand. Consider a stand of pine which started out as a plantation 30 years ago. Since all trees will be the same age, you might expect that they will all be the same diameter 30 years after planting. This is not so, for the reason that some trees in the stand will have achieved a dominant position early in life, while others have become intermediate and still others will be suppressed. The taller trees will be more vigorous in all respects, including diameter growth; the opposite will be true for the intermediate and suppressed trees. If all the trees in the stand were measured for diameter at the same time, it would be found that most of them would be close to the mean diameter for the whole stand, while some would be considerably larger or smaller than the average. A tally of diameters will show that the diameter class containing the mean value will have the greatest frequency and that the frequency of those above or below the mean will be progressively less.

2. distribution of heights of male students at a University. We have already discussed the heights of University students in connection with obtaining the mean height, but let us look at the distribution of heights. The occurrence of men 6'11" tall would be very infrequent; likewise there would not be many who were 5'0". The mean height of male University students at the present time is approximately 5'8½". If a large sample, say 1000 or more students, were obtained, the distribution of heights would approximate a normal curve.

An empirical approximation of a normal curve

An approximation of a normal curve can be built up by tossing 10 coins at one time and counting the number of heads turned up in the 10 coins. It is possible in a single toss to get 10 heads or 10 tails (no heads), but the chances of either of these occurring is very slight. It is more likely that some number of heads close to the mean number of heads (5 heads) will turn up.

As described above, this type of distribution is really a binomial distribution since the X variable (number of heads) is discrete, but it was also stated that as the number of tosses increases, the curve will approach that of a normal curve.

An experiment was conducted to demonstrate this situation. Ten coins were thoroughly shaken and tossed in a group; for each toss, the number of heads showing was tallied. There was no control over the coins and they were allowed to fall completely at random. If one coin in the 10 struck an object on the desk or rolled off, the toss was cancelled in order to achieve as random a sample as possible. Of course, we assume in this type of experiment that the coins are all identical in shape and weight and that the method of tossing did not introduce any bias. Figure 4.2 shows the frequency polygons after 125, 250, 500 and 1000 tosses were made. As the number of tosses increased, the distribution came closer and closer to a normal distribution.

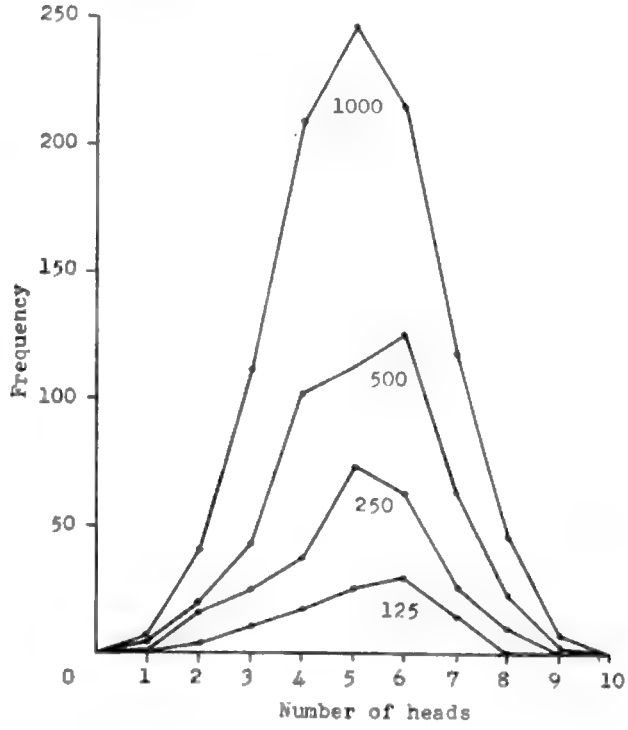


Figure 4.2 - Frequency polygons of 125, 250, 500 and 1000 tosses of 10 coins at a time.

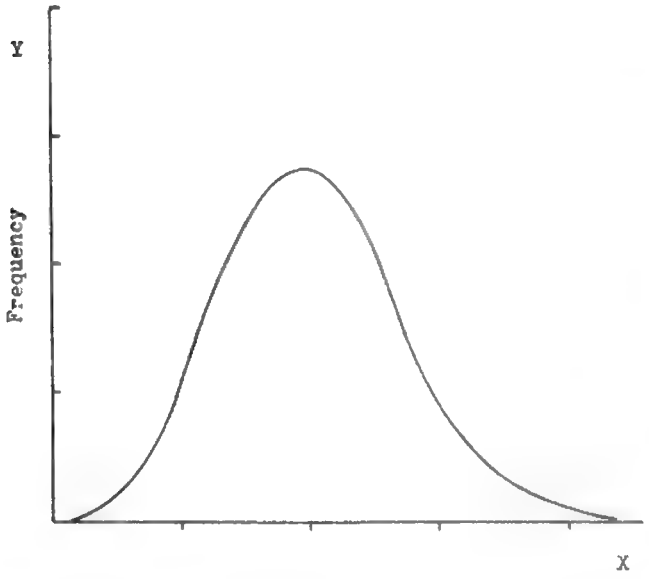


Figure 4.3 - Right- or positively-skewed curve.

which are very large or small respectively. Figures 4.3 and 4.4 illustrate the right- or positively-skewed distribution and the left- or negatively-skewed distribution. For the mathematician, there are methods of determining the amount of skewness, but in this beginning treatment of statistics, we are concerned only with being able to recognize and describe a departure from a normal curve.

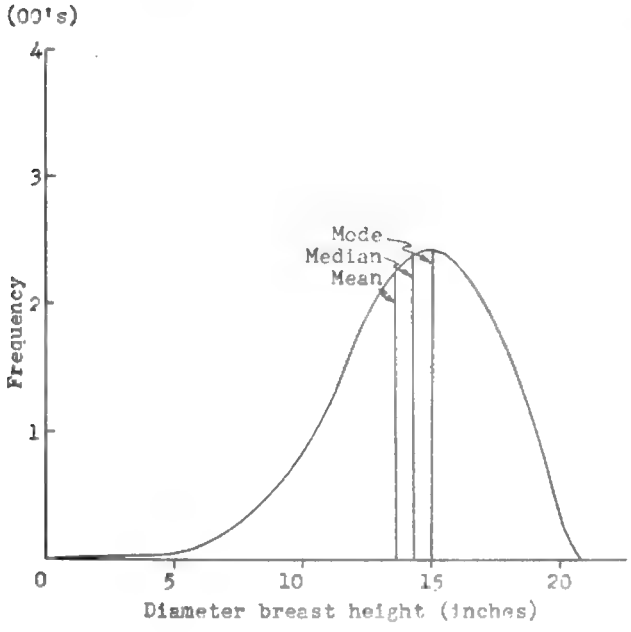


Figure 4.5 - Left skewed curve showing position of the mean, median and mode.

Non-normal curves

Since it is rare to find a perfectly normal distribution in nature, there are special terms to describe distributions which are nearly normal; the terms are "skewness" and "kurtosis" and will be illustrated in the following sections.

Skewness

A distribution that is skewed has the general shape of a normal curve except that either the right or left tail of the curve is elongated. This is caused by a small number of X values

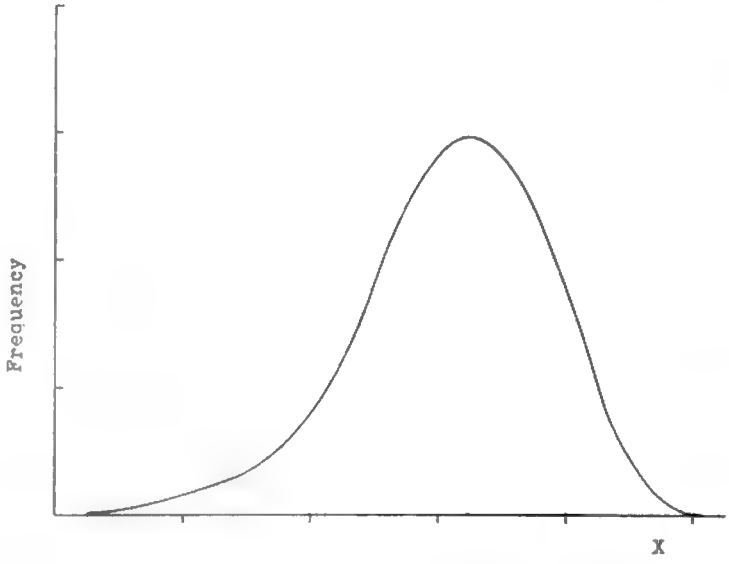


Figure 4.4 - Left- or negatively-skewed curve.

In a skewed distribution, the value of the mean, median and mode are not identical as in a normal distribution. The value of the mean is affected by the size and number of the items, so will be pulled toward the side on which the longer tail occurs. The median is affected by the number of items only, not by their size, and will occupy a position between the mean and the mode. In all skewed distributions, the order of these three statistics, going toward the mode from the long tail, is mean - median - mode. A skewed distribution showing these points is given in Figure 4.5.

Kurtosis

This term is given to the general characteristic of "peakedness". Not all frequency distributions follow the normal pattern; some have an abnormally large number of items at or close to the mean value while others have frequencies resulting in a flattened curve. The first is called a "lepto-kurtic" curve while the second is a "plati-kurtic" curve. These two, in relation to a normal distribution, are illustrated in Figure 4.6.

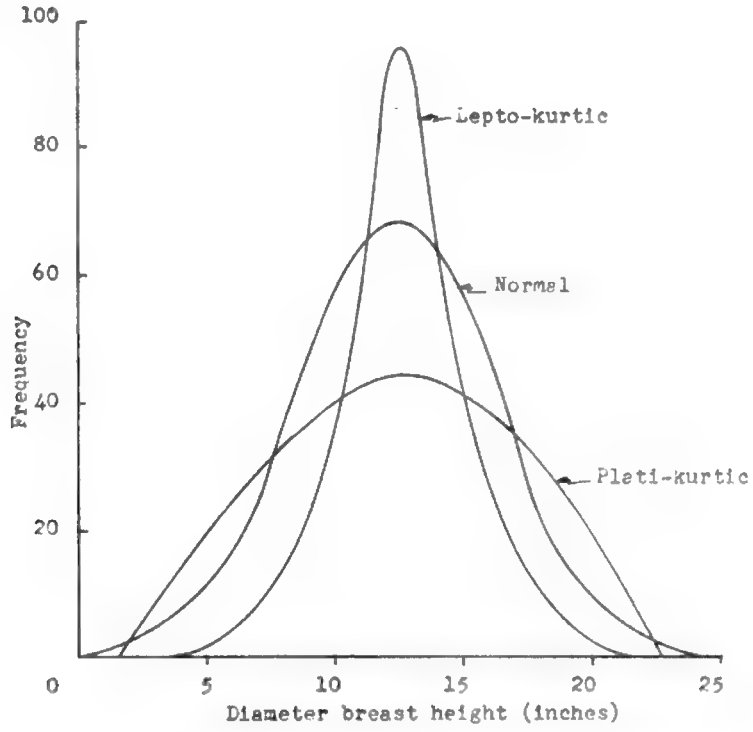


Figure 4.6 - Comparison between a normal, lepto-kurtic and plati-kurtic curve.

Chapter 5

MEASURES OF DISPERSION

The comparison between two stands on the basis of mean diameter alone does not provide sufficient information. Mean diameter does not tell anything about the manner in which the sampling units are distributed about the mean. In statistical work, the distribution of the items is as important as the mean or other statistics.

To illustrate the importance of distribution, consider two different stands in which the mean diameters are exactly the same. Table 5.1 shows Stand 1 having a mean of 13.68 inches and a range of diameters from 10 to 22 inches inclusive; Stand 2 has the same diameter but a range from 11 to 16 inches inclusive. Both stands have the same number of trees. These two stands represent completely different situations. Consider each in terms of a certain product requiring a mean diameter of 14 inches; Stand 1 would not be thought of as having a good distribution for the particular product while Stand 2 does.

Table 5.1 - Diameter distribution of two stands having the same number of trees and the same diameter.

dbh	Number of trees per acre		Calculations for mean	
inches	Stand 1	Stand 2	fX1	fX2
10	2	0	20	0
11	3	2	33	22
12	5	3	60	36
13	2	4	26	52
14	2	3	28	42
15	1	4	15	60
16	1	3	16	48
17	0	0	0	0
18	0	0	0	0
19	0	0	0	0
20	2	0	40	0
21	0	0	0	0
22	1	0	22	0
	19	19	260	260
			$\bar{X}$ 13.68	13.68

As this illustration shows, the mean does not present the whole picture. We would be interested in knowing, in any distribution, how the items are grouped, whether they are distributed somewhat symmetrically about the mean or not and the total range of values from the highest to the lowest.

To answer these questions, we turn to measures of dispersion; the first, and probably the most important is -

Standard deviation

A measure of dispersion which determines the amount of variability of the sampling units

about the mean value is called standard deviation. For ungrouped data, it has the formula: -

$$s = \sqrt{\frac{\Sigma x^2}{N - 1}} \qquad (5.1)$$

where s = standard deviation

Σx^2 = sum of the squared deviations of each item from the mean $(X - \bar{X})$

$N - 1$ = degrees of freedom.

Notice that each deviation from the mean is squared and the sum of the squared values is used as the numerator in the equation.

N - 1 is used as the denominator rather than N, the total number of items; this is a statistical device to make sure that unbiased estimates of population parameters will be obtained. You will meet and use degrees of freedom constantly in further statistical work.

Standard deviation is a very useful tool and is one of the most widely used terms in the field of statistics. A number of formulas and statistical tests require that the value of the standard deviation be known; it is a cornerstone upon which much statistical inference is built.

Let us work out an example to show the steps and mathematics of calculating standard deviation. Having obtained the mean diameter of the distribution, the steps are: -

1. for each item in the list, obtain the deviation $X - \bar{X}$ symbolized by x. KEEP THE ALGEBRAIC SIGN UNLESS THE DEVIATION IS ZERO.
2. square each deviation to obtain x^2.
3. sum the squared deviations to obtain Σx^2.
4. substitute Σx^2 and N - 1 in the standard deviation formula and solve. The answer can be plus or minus and is written ± s.

Table 5.2 - Calculation of standard deviation.

dbh inches	x inches	x^2 inches
10	- 3.5	12.25
10	- 3.5	12.25
10	- 3.5	12.25
11	- 2.5	6.25
11	- 2.5	6.25
12	- 1.5	2.25
12	- 1.5	2.25
12	- 1.5	2.25
12	- 1.5	2.25
13	- 0.5	0.25
14	0.5	0.25
14	0.5	0.25
15	1.5	2.25
15	1.5	2.25
15	1.5	2.25
15	1.5	2.25
16	2.5	6.25
17	3.5	12.25
18	4.5	20.25
18	4.5	20.25
Sums: $\Sigma X = 270$	$\Sigma x = 0$	$\Sigma x^2 = 127.00$

$$\bar{X} = \frac{270}{20} = 13.5$$

(Continued on next page)

$$s = \sqrt{\frac{127.00}{20 - 1}}$$

$$= \sqrt{\frac{127.00}{19}}$$

$$= \sqrt{6.684}$$

$$= \pm 2.58 \text{ inches}$$

We now have a figure ± 2.58 inches; what does it mean? What can we interpret from it and how does it tell us the variation in the stand? Remember we are looking for a measure of dispersion and standard deviation is perhaps the most important of all methods. "Properties of a Normal Curve" will be discussed shortly; until we reach that section, it is sufficient to say that approximately 68% of the trees measured for the sample will be included in the range $\bar{X} \pm s$. If the sample is a representation of the population from which it was drawn, we may deduce that the population will have a similar variation.

Before looking at the properties of a normal curve, there are two other methods of calculating standard deviation, one depending upon the way in which the data are listed and the other on the actual amount of data.

Standard deviation for grouped data

The previous example used the raw data or individual measurements and you will admit that the calculation of standard deviation is somewhat tedious. With a large number of items and with the deviations being taken to two places of decimals, the possibilities of computational errors are large. Just as the mean could be calculated by grouping the items into diameter classes, the standard deviation can be calculated similarly. We merely group the data into convenient classes and apply group frequencies throughout. The same data as before is presented in Table 5.3 except that the data have been grouped into diameter classes.

Table 5.3 - Calculation of standard deviation using grouped data.

dbh inches	Number of trees f	x inches	x^2 inches	fx^2
10	3	- 3.5	12.25	36.75
11	2	- 2.5	6.25	12.50
12	4	- 1.5	2.25	9.00
13	1	- 0.5	0.25	0.25
14	2	0.5	0.25	0.50
15	4	1.5	2.25	9.00
16	1	2.5	6.25	6.25
17	1	3.5	12.25	12.25
18	2	4.5	20.25	10.50
Sums:	20	0		127.00

Mean: 13.5 inches

$$= \sqrt{\frac{\Sigma fx^2}{\Sigma f - 1}} \quad * \qquad (5.2)$$

$$= \sqrt{\frac{127.00}{19}}$$

$$= \sqrt{6.684}$$

$$= \pm 2.58 \text{ inches}$$

* For grouped data, N is the equivalent of Σf, the sum of all frequencies. In ungrouped data, N represents the number of X classes but is also equivalent to Σf, because each X class has a frequency of 1.

Short-cut method of calculating standard deviation

The two methods described so far are the long methods of obtaining standard deviation. While they may not seem too laborious for a small number of items, the task becomes overwhelming when the number of classes is large or when the mean value of X is a decimal. The latter is true more often than not. There is a short-cut method which is an extension of the method we used for determining the mean from an assumed mean. To refresh your memories, here is the formula we used to obtain the mean: -

$$\bar{X} = \bar{X}_A + \frac{\Sigma fx'}{\Sigma f}$$

The formula for determining the standard deviation from an assumed mean is: -

$$s = \sqrt{\frac{\Sigma fx'^2 - \frac{(\Sigma fx')^2}{\Sigma f}}{\Sigma f - 1}} \qquad (5.3)$$

The symbols are the same as we used previously. The saving of computational labor is considerable when both the mean and standard deviation can be calculated simultaneously. Since we are now taking deviations from an assumed mean rather than from the true mean, the deviations are symbolized as x'; the sum of the squared deviations will be too large or too small depending on whether the assumed mean is larger or smaller than the true mean. You will realize too, that the second term in the numerator in Equation 5.3 is the correction factor which takes into account the difference between the assumed and the true mean of the distribution. These comments are only for the purpose of explaining the terms in Equation 5.3. In actual practice, we do not determine the true mean ahead of time to see how far the assumed mean deviates from it; this would defeat the purpose of the formula. We assume a mean, determine deviations from the assumed mean and then correct the sum of squares based on these deviations from the assumed mean.

The advantages of determining the standard deviation by the assumed mean method are (1) the mean and standard deviation can be calculated simultaneously (2) the deviations, and hence the deviations squared, can be determined without introducing decimals in the calculation.

The same set of data as we had previously will be used to determine standard deviation by the short-cut method.

Table 5.4 - Calculation of standard deviation by the short-cut method.

dbh inches	Number of trees f	x'	fx'	fx'^2
10	3	- 4	- 12	48
11	2	- 3	- 6	18
12	4	- 2	- 8	16
13	1	- 1	- 1	1
14	2	0	0	0
15	4	1	4	4
16	1	2	2	4
17	1	3	3	9
18	2	4	8	32
Sums:	$\Sigma f = 20$		$\Sigma fx' = -10$	$\Sigma fx'^2 = 132$

True mean: $\bar{X}_A = 14$

$$\bar{X} = 14 + \frac{(-10)}{20}$$

$$= 14 - 0.5$$

$$= 13.5$$

(Continued on next page)

Standard deviation:

$$s = \sqrt{\frac{132 - \frac{(-10)^2}{20}}{20 - 1}}$$

$$= \sqrt{\frac{132 - \frac{100}{20}}{19}}$$

$$= \sqrt{\frac{132 - 5}{19}}$$

$$= \sqrt{\frac{127}{19}}$$

$$= \sqrt{6.684}$$

$$= \pm 2.58 \text{ inches}$$

The answer is exactly the same as we had before but the computations are much more simple. It makes no difference what value of $\bar{X}_A$ we choose; the correction factor will take care of the difference between it and the true mean.

Range as an estimate of population standard deviation

A sample is taken for one purpose, to obtain a good estimate of population parameters. We are not primarily interested in the sample per se, but with the population from which the sample is chosen.

The sample mean ($\bar{X}$) serves as an estimate of the population mean (μ)* while the sample standard deviation (s) gives an estimate of the population standard deviation (σ). The numerical value of standard deviation depends upon (1) the number of items in a sample and (2) the range in value from the highest to the lowest. This latter is an indication of the variability exhibited by the sample. In the following section, we are assuming that we are dealing with a normally-distributed population from which an unbiased sample is drawn. In actual practice, a sample may be drawn from a population having a lepto-kurtic distribution which will result in a low value of s, or a plati-kurtic one in which the value of s is high. Table 5.5[1] shows the relative efficiency using range as an estimate of σ and the ratio of s to the range for samples of different sizes drawn from a normally-distributed population.

Table 5.5 - Ratio of s to Range, and Relative Efficiency of Samples of Different Sizes.

n	$\frac{s}{\text{Range}}$	Relative Efficiency	n	$\frac{s}{\text{Range}}$	Relative Efficiency
2	0.886	1.000	12	0.307	0.815
3	0.591	0.992	14	0.294	0.783
4	0.486	0.975	16	0.283	0.753
5	0.430	0.955	18	0.275	0.726
6	0.395	0.933	20	0.268	0.700
7	0.370	0.912	30	0.245	0.604
8	0.351	0.890	40	0.231	0.536
9	0.337	0.869	50	0.222	0.490
10	0.325	0.850			

For example, 20 items in a sample having a total range of 15 units, would have a population standard deviation of (15)(0.268) = ± 4.02 units. The third column gives the relative

* Greek lower case letters are used for the population parameters, while Arabic letters refer to the sample statistics.

[1] Reproduced by permission of George W. Snedecor: STATISTICAL METHODS (5th Edition 1956), copyright, Iowa State University Press, Ames, Iowa.

efficiency of estimating σ as against calculating s from the sample. In the example given, the relative efficiency is 0.700, meaning that 20/0.700 = 28.5 items would give as accurate an estimate of σ using the range as 20 items would if the standard deviation were actually calculated. Estimating σ by using the range becomes an economic problem at times. If the units in the problem were cords per acre measured on fifth-acre plots, it is immediately apparent that it is more economical to calculate s from the sample than to measure an additional 8 or 9 plots to obtain the same accuracy of estimate.

Standard deviation and the normal curve

In addition to having the three characteristics stated in Chapter 4, a normal curve has specific relationships between the value of the mean and the standard deviation. If we were to draw a perfectly normal curve with the mean, median and mode identical in value, the area under the curve between X - s and X + s is 68.26% of the total area under the curve. The area under the curve between X - 2s and X + 2s is 95.45% and between X - 3s and X + 3s is 99.73%. Where do these figures come from? How can we state them so exactly? They are derived from the formula for the normal curve and as such, are exact percentages.

The "tails" of the curve theoretically go out to infinity since, in an infinite-sized population, there are always a chance of a value infinitely small or infinitely large occurring. However, for convenience, the upper and lower limits of the population are set at -3s and +3s.

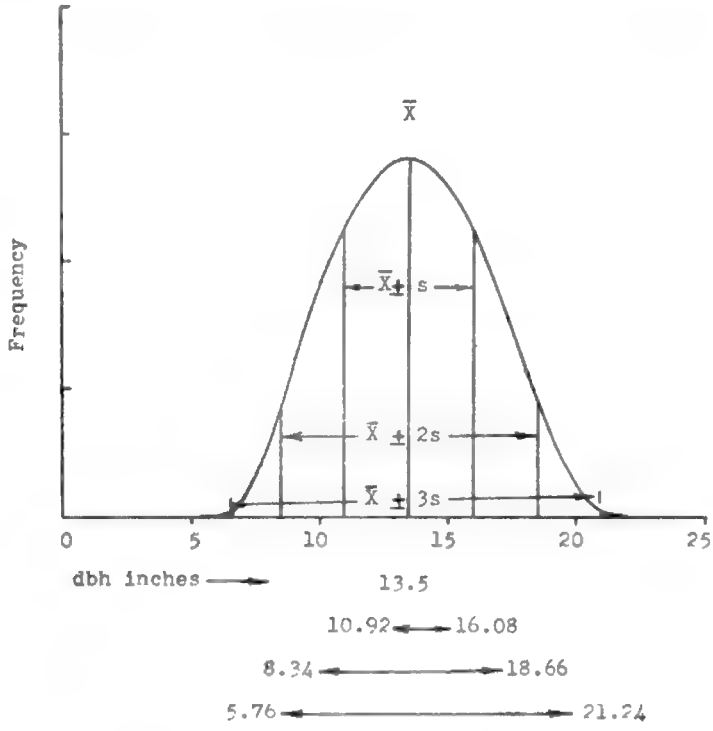


Figure 5.1 - Normal curve with $\bar{X}$ = 13.5 inches, s = ± 2.58 inches and limits of ± s, ± 2s and ± 3s.

Figure 5.1 shows a normal curve with a mean of 13.5 inches and a standard deviation of ±2.58 inches with the limits of ±s, ±2s and ±3s. We generally refer to these limits as "fiducial limits".

Normal deviate or Z

Sometimes it is more convenient to express deviations from the mean in terms of normal deviates or standard units, given the symbol Z. In the sample of diameters in which the mean was 13.5 inches and standard deviation was ±2.58 inches, a deviation $(X - \bar{X})$ of 2.58 inches divided by the standard deviation would result in a normal deviate of 1.0. The formula for normal deviate is: -

$$Z = \frac{x-\mu}{\sigma} \qquad \text{or} \qquad \frac{x}{s} \qquad . \quad (5.4)$$

In a perfectly normal curve, the area under the curve between the mean (maximum ordinate) and a particular value of Z can be determined very precisely. The Z notation has the advantage of universality in that we are dealing with populations having different standard deviations but which are expressed in standard units. Thus, the particular value of s is of no consequence as long as we express a deviation (x) in terms of the number of standard deviations.

The normal deviate has an interesting application in that it can be used to determine the probability of occurrence of a specific value of X. You are familiar with the term "probability" as it is used in every-day conversation, although the statistical meaning may not be clear. It is nothing more than expressing the odds or chances of a certain event occurring. In a normal distribution, we would not expect a value of X which is far removed from the mean

to occur very often. Let us assume that a normally-distributed population has a population mean of 20.8 pounds and a population standard deviation of ± 6.4 pounds. An X value of 17.6 pounds might be expected to occur quite often since it is close to the mean, while an X of 8.0 pounds would seldom occur. We have used the descriptive terms "quite often" and "seldom", but in statistics, we must be more precise. We would like to attach a numerical value to "often" and "seldom". Let us examine the two X values in terms of deviations and, more particularly, of standard units.

$X = 17.6$ pounds	$X = 8.0$ pounds
$X - \mu = 17.6 - 20.8$	$X - \mu = 8.0 - 20.8$
$= -3.2$ pounds	$= -12.8$ pounds
$Z = \frac{-3.2}{6.4}$	$Z = \frac{-12.8}{6.4}$
$= -0.5$	$= -2.0$

We are now ready to look at Table 5.6 which gives the area under a normal curve between the maximum ordinate and the ordinate at 0.1 intervals of Z. Table A.2 in the Appendix shows the area under a normal curve bounded by the mean and values of Z in .01 intervals.

Returning to the problem, we find that the areas under the normal curve are: -

X	Z	Area under the curve
17.6	- 0.5	.19146
8.0	- 2.0	.47725

For each value, we can ask "What percent of the items making the population is smaller than the given X value?" Remember that the whole population is represented by the full curve, while the table of Z values is for one-half the curve only. There X = 17.6 pounds with Z = 0.5, the percent of items smaller in value is .50000 - .19146 = .30854, or 30.854%.

Table 5.6 - Areas under a normal curve between the maximum ordinate and the ordinate for 0.1 intervals of Z.

Z	Area	Z	Area
0.0	.00000	2.0	.47725
0.1	.03983	2.1	.48214
0.2	.07926	2.2	.48610
0.3	.11791	2.3	.48928
0.4	.15542	2.4	.49180
0.5	.19146	2.5	.49379
0.6	.22575	2.6	.49534
0.7	.25804	2.7	.49653
0.8	.28814	2.8	.49744
0.9	.31594	2.9	.49813
1.0	.34134	3.0	.49865
1.1	.36433	3.1	.49903
1.2	.38493	3.2	.49931
1.3	.40302	3.3	.49952
1.4	.41924	3.4	.49966
1.5	.43319	3.5	.49977
1.6	.44520	3.6	.49984
1.7	.45543	3.7	.49989
1.8	.46407	3.8	.49993
1.9	.47128	3.9	.49995

For X = 8.0 pounds with Z = - 2.0, the percent is .50000 - .47725 = .02275 or 2.275%.

Another problem which arises in the use of Z is to determine the percent of items occurring between various values of X. For instance, what percent would occur between X = 11.2 pounds and X = 27.2 pounds? Transform each X into Z as follows: -

X = 11.2 pounds	X = 27.2 pounds
$X - \mu$ = 11.2 - 20.8	$X - \mu$ = 27.2 - 20.8
= - 9.6 pounds	= 6.4 pounds
$Z = \frac{-9.6}{6.4}$	$Z = \frac{6.4}{6.4}$
= - 1.5	= 1.0
Area from Table = .43319	= .34134

Add the two areas, .43319 + .34134 = .77453 and we conclude that 77.453% of the items in the population will lie between the limits of 11.2 pounds and 27.2 pounds.

You should become familiar with three values derived from Table 5.6 as you will meet them frequently in statistical work. They are: -

$$\bar{X} \pm s = 68.268\%$$

$$\bar{X} \pm 2s = 95.450\%$$

$$\bar{X} \pm 3s = 99.730\%$$

Probability

If an event occurs one time in a hundred trials, it is said to occur with a probability of 1 in 100 or 1:100. An event occurring five times in a hundred trials has a probability of 1:20. Looking back at the limits given on page 31, we can express the probabilities of occurrence as: -

Limits	Exact Probability	Approximate Probability
$X \pm s$	.68268	2:3
$X \pm 2s$	.95450	19:20
$X \pm 3s$	.99730	99:100

These probabilities are approximate only, but serve as a convenient reference point in discussing limits. Consider a population of trees having a mean diameter of 15.0" and a standard deviation of ± 3.0"; if we state that the probability of the number of trees between the limits of 12.0" and 18.0" is 0.68268, we can surmise that approximately two-thirds of the trees are in this range of diameters. Also, we know that we are dealing with deviations from the mean ($X - \bar{X}$) equal to - 3.0" and + 3.0" and that these are the same value as the standard deviation. Since $X - \mu$ ± 3.0" and s = ± 3.0", we have a Z of ± 1.0. The probability of .68268 came from Table 5.6 for Z = 1.0. The tabular value was doubled because we are considering deviations on both sides of the mean rather than on just one.

We do not need to restrict our thinking to deviations from the mean that are exact multiples of the standard deviation, although it is convenient to do so. The table of Z values given either on page 32 or in the Appendix will establish probabilities for any value of $X - \mu$ providing it does not exceed Z = 3.99. The normal curve is such that very little accuracy is sacrificed beyond Z = 3.0; only 0.27% of the total area under the curve is included beyond the limits of Z = ± 3.0.

Coefficient of variation

An interesting application of standard deviation is the comparison between two entirely different types of measurements. Knowing that one sample had a deviation of ± 6.6 feet and another ± 4.2 inches would not enable us to tell which sample showed the greater variability. Instead of comparing the standard deviations, we turn to a comparison of their relative dispersions. We are interested in the standard deviation, of course, but now think of it in relation to the mean. The formula for the coefficient of variation is: -

$$V = \frac{s}{\bar{X}} (100) \qquad (5.5)$$

where V = coefficient of variation

s = standard deviation

$\bar{X}$ = mean of the sample.

Coefficient of variation is a number without units of measurement since both the numerator and denominator are in the same units. Multiplying the fraction by 100 establishes the coefficient of variation as a percentage.

Two examples of the use of coefficient of variation are given here to acquaint you with the use of this statistic and its interpretation.

Example 1

A sample of trees was taken from each of two stands which differed widely in age. The object was to determine whether the young stand was more variable in its diameter distribution than the older one. The answer should be obvious, but the computation will be carried through to illustrate the use of coefficient of variation. The statistics for each sample are as follows: -

		Stand 1	Stand 2
Average age		20 years	60 years
Mean diameter	$\bar{X}$	4.5 inches	9.7 inches
Standard deviation	s	± 0.8 inches	± 2.7 inches
Coefficient of variation	V	$= \frac{0.8(100)}{4.5}$	$= \frac{2.7(100)}{9.7}$
		= 17.7%	= 27.8%

We can conclude that the relative dispersion of diameters for the 60-year old stand is 10% greater than for the 20-year old one.

Example 2

A plantation is now 15 years old. The site, based on soil, topography and other physiographic factors, indicates that growth is better on a north slope than on a south slope. In addition to diameter growth, the height growth is also an indication of a plant's response to its environment. Is the height growth more variable on the north slope than on the south? A sample of heights is taken from each site condition with the results: -

	North slope	South slope
Mean height growth per year	8.8 inches	5.3 inches
Standard deviation	± 2.20 inches	± 2.12 inches
Coefficient of variation	$= \frac{2.20(100)}{8.8}$	$= \frac{2.12(100)}{5.3}$
	= 25.0%	= 40.0%

We conclude that the variation in height growth is greater on south slopes than it is on north slopes for the species in question and on the sites measured. You can not generalize from the results of two samples and say that this would be true of all species on all sites.

Rejection of abnormal data

Quite frequently in forest samples, we run across one or two measurements which seem to be abnormally high or abnormally low; they do not conform to the general trend of measurements for the rest of the sample. Are these seemingly abnormal measurements to be included in the calculation of the mean, standard deviation and other statistics, or can we reject them as being non-representative of the population being sampled? Another way of saying the same question is to ask whether these seemingly abnormal measurements might not be from a different population from the one in which we are interested. The inclusion of a very high or low value has an undue influence on the value of the standard deviation since the deviation $(X - \bar{X})$ must be squared.

Before going further on this subject, it must be pointed out that rejection of abnormal data is not according to accepted statistical practice which assumes that we are sampling an infinite-sized population in which any measurement is theoretically possible.

It is usual in forestry practice to reject those measurements which have deviations from the mean which are equal to or are greater than 2.5 standard units, or where $Z \geq 2.5$[1/] This represents a probability of 1.24 in 100 of an item as large or larger (or as small) occurring again in a sample of the same size taken from the same population.

An instance sometimes occurs in forestry measurements where we wish to determine the mean and standard deviation of ages of trees in what we believe to be an even-aged stand. Basically, the population which we wish to sample is that of even-aged trees, but we do not know the ages until we have obtained cores with an increment borer. Our basis for sampling is not tree age, but height or some other characteristic which is related to age. The sample, then, is taken from dominant and codominant trees in the stand. Unfortunately, height is related not only to age but also to stand conditions which have existed in the past, and we could obtain ages from trees of approximately the same height which indicate that some of the trees might have been suppressed in their early stages. Let us examine such a situation.

Sixteen trees were bored for total age determination and the sample was selected from dominant and codominant trees in a stand. The ages and the calculations of the mean and standard deviation are included in Table 5.7.

Table 5.7 - Calculation of mean and standard deviation of 16 representative trees from an even-aged stand.

Tree Number	Total age years	x'	x'^2
1	58	- 2	4
2	62	+ 2	4
3	65	+ 5	25
4	57	- 3	9
5	66	+ 6	36
6	64	+ 4	16
7	59	- 1	1
8	55	- 5	25
9	90	+30	900
10	71	+11	121
11	65	+ 5	25
12	63	+ 3	9
13	68	+ 8	64
14	67	+ 7	49
15	72	+12	144
16	60	0	0
$\bar{X}_A = 60$		$\Sigma+ = +93$	1432
		$\Sigma- = -11$	
		$\Sigma x' = +82$	

(Continued on next page)

1/ The symbol $\geq$ is read "equal to or greater than."

$$\bar{X} = 60 + \frac{82}{16}$$

$$= 60 + 5.1$$

$$= 65.1$$

$$s = \sqrt{\frac{1432 - \frac{(82)^2}{16}}{15}}$$

$$= \sqrt{\frac{1432 - 420.2}{15}}$$

$$= \sqrt{\frac{1011.8}{15}}$$

$$= \sqrt{67.45}$$

$$= \pm 8.21 \text{ years}$$

On the basis of our standards of rejection, any item which is equal to or exceeds $\bar{z} = 2.5$ should be rejected as being abnormal. In this instance, the limit of rejection would be any item which has an $X - \bar{X}$ equal to or greater than $2.5(8.21) = 20.62$ years. Check the large deviations in the x' column, but remember that these are deviations from the assumed mean, not the true mean of the sample. We have three deviations which are over 10 in value - Tree numbers 9, 10 and 15. Even allowing for the fact that the true mean is 65.1 years, would not eliminate Tree numbers 10 and 15 from the sample. Look at Tree number 9; the deviation from the true mean is 90 - 65.1 = 24.9 years which exceeds the limit of $\bar{z} = 2.5$, so we reject it as being abnormal. If an item is rejected, the remaining items in the sample must be re-calculated for the mean and standard deviation.

$$\bar{X} = 60 + \frac{52}{15}$$

$$= 60 + 3.4$$

$$= 63.4$$

$$\Sigma+ = + 93$$

$$\Sigma- = - 11$$

$$\Sigma x' = + 82$$

$$\text{Reject } x' \text{ of } 30 \quad - 30$$

$$\text{Revised } \Sigma x' = + 52$$

$$\Sigma x'^2 = 1432$$

$$\text{Reject } x'^2 \text{ of } 900 = -900$$

$$\text{Revised } \Sigma x'^2 = +532$$

$$s = \sqrt{\frac{532 - \frac{(52)^2}{15}}{14}}$$

$$= \sqrt{\frac{532 - 180.2}{14}}$$

$$= \sqrt{\frac{351.8}{14}}$$

$$= \sqrt{25.1}$$

$$= \pm 5.09 \text{ years}$$

The new limits of rejection now are $2.5(\pm 5.09)$ years = $\pm$ 12.72 years. The new $\bar{X}$ is 63.4 years, so any item greater than 63.4 + 12.7 = 76.1 years, or less than 63.4 - 12.7 = 50.7 years would be rejected on a subsequent calculation. There are no items in the revised list which exceed these limits, so we can conclude that the tree having an age of 90 years was abnormal and that the remainder belonged to the population in which we were interested.

Chapter 6

STANDARD ERROR

Standard error of the mean

In any sample taken from an infinite population, we expect that the items in the sample will vary from the mean; the amount that they vary will determine the value of the standard deviation. The sample standard deviation (s) is an estimate of the population standard deviation (σ). In addition to an estimate of σ, we are interested in estimating the value of the population mean (μ); standard deviation does not tell us this, and nothing is said or implied about the population mean. If we took a 100% tally of all items in an infinitely large population, we could obtain a precise population mean; this is obviously impossible and even for large but finite populations, impractical.

Consider a population consisting of even-aged white oak trees covering an area of approximately 1000 acres. The number of trees in this population is obviously large but not infinite; it might number 30,000 or 200,000 depending upon the age of the trees. Now let us take an unbiased, representative sample from the population. The sample will give us a sample mean ($\bar{X}$) and a sample standard deviation (s). Let us take another sample. Will the second one have the same mean and standard deviation as the first? It is possible but unlikely that the two sets of statistics will be identical. We could continue sampling from this population until we had a large number of samples, each with its own mean and standard deviation. Taking the sample means, we could group them into convenient classes and arrange them in order by frequency. Providing that enough samples were taken, the frequency distribution of sample means will resemble a normal distribution curve as shown in Figure 6.1

Note that the abscissa is graduated into units of mean values of X. The mean of this distribution is not the familiar $\bar{X}$, but the average of a number of means which is called the mean of means and symbolized as $\bar{X}_{\bar{X}}$.

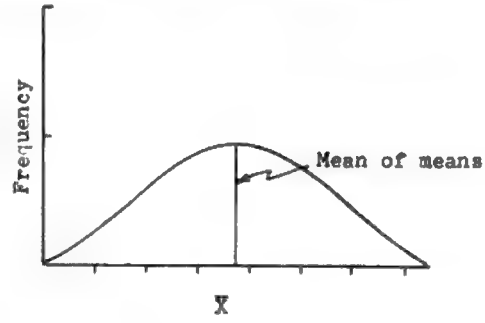


Figure 6.1 - Distribution of the means of a number of samples.

Unless we took an infinite number of samples, we could not determine the value of the mean of means, but since the distribution curve in Figure 6.1 is made up of a large number of samples, we can rely on measures of central tendency and say that the population mean will lie close to the mean of means. How close, we do not know at the present time; it depends upon the number of samples taken and the variation between sample means.

Remember that the primary purpose of taking a sample is to estimate population parameters. It would be a long expensive process to estimate the population mean by securing a large number of sample means. Fortunately, there is a formula which allows us to estimate the range in which the population mean will lie by using the statistics from a single sample. Naturally, the sample we use must consist of unbiased measurements and the number of items in the sample must be large enough to be representative of the population. The formula is: -

$$s_{\bar{X}} = \frac{s}{\sqrt{\Sigma f}} \quad \text{or} \quad \frac{s}{\sqrt{N}} \tag{6.1}$$

where $s_{\bar{X}}$ = standard error of the mean

s = standard deviation of the sample

N or Σf = number of items in the sample.

Since the curve in Figure 6.1 is a normal curve, the same constants apply as in our previous study of areas under the normal curve and probabilities. Therefore, we can say that the population mean (μ) will lie within the range of the sample mean $\pm$ the standard error with a probability of 2:1. Putting this into a symbolized statement, we have: -

$$\bar{X} - s_{\bar{x}} < \mu < \bar{X} + s_{\bar{x}} \qquad (6.2)$$

which states that the population mean is between the numerical value of $\bar{X} - s_{\bar{x}}$ and $\bar{X} + s_{\bar{x}}$. The following statements are also true: -

Fiducial Limits	Approximate Probability	Exact %
$\bar{X} - s_{\bar{x}} < \mu < \bar{X} + s_{\bar{x}}$	2:1	68.26
$\bar{X} - 2s_{\bar{x}} < \mu < \bar{X} + 2s_{\bar{x}}$	19:1	95.45
$\bar{X} - 3s_{\bar{x}} < \mu < \bar{X} + 3s_{\bar{x}}$	99:1	99.73

The following example will illustrate the method of estimating the population mean from a single sample.

A sample consisting of 121 diameters from an even-aged stand resulted in a mean of 7.8 inches and a standard deviation of $\pm$ 5.6 inches. Our problem is to determine within what limits the population mean will lie at the probability levels of 2:1, 19:1 and 99:1. Substituting the sample statistics into Formula 6.1, we have: -

$$s_{\bar{x}} = \frac{5.6}{\sqrt{121}}$$

$$= \frac{5.6}{11}$$

$$= \pm 0.514 \text{ inches}$$

According to our fiducial limits, the population will lie within the following: -

1. probability of 2:1

$$7.8 \pm 0.514 = 7.286 \text{ to } 8.314 \text{ inches}$$

2. probability of 19:1

$$7.8 \pm 2(0.514) = 6.772 \text{ to } 8.828 \text{ inches}$$

3. probability of 99:1

$$7.8 \pm 3(0.514) = 6.258 \text{ to } 9.342 \text{ inches}$$

In this example, there is a wide variation in the original sampling units; the sample size is adequate (121), so we must conclude that the population itself is quite varied. If the sample of 121 trees was obtained without bias and is a true representation of the population, we can accept the sample statistics without question. As the variation in the original sample increases, we must expect our estimate of the population mean to be within an increasingly wide range, since s controls the size of $s_{\bar{x}}$, providing the size of the sample does not change.

The statistic 't'

The statistic 't' is vitally concerned with fiducial limits (sometimes called interval estimates) and tests of hypotheses. The distribution of 't' was first discovered by W. S. Gossett in 1908; he wrote under the pseudonym of "Student" and the statistic is often referred to as Student's 't' distribution. R. A. Fisher did some further work on the 't' distribution in 1924 and brought it to its present degree of completion.

The quantity 't' has the formula: -

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}} \tag{6.3}$$

where $\bar{X}$ = mean of the sample

s = standard deviation of the sample

$s_{\bar{x}}$ = standard error of the sample

μ = population mean.

You realize that the quantity $s_{\bar{x}}$ may be replaced by its own formula $\frac{s}{\sqrt{\Sigma f}}$ or $\frac{s}{\sqrt{N}}$.

The statistic 't' is a measure of the deviation of a sample mean from the population mean in terms of the number of standard errors. The distribution of 't' is practically normal for large samples, being symmetrical about the mean. For small samples, where N = 30 or less, 't' plays an increasingly important role in estimating fiducial limits. Instead of having approximate probabilities, we are now in a position to state exact probabilities, of which 5% and 1% are the most commonly applied. A probability of 5% implies that we will be right in our estimate 95 times out of 100; another way of saying it is that our estimate will be correct unless a 1 in 20 chance has occurred.

We might choose the 5% level of probability and express 't' as: -

$$-t_{.05} < \frac{\bar{X} - \mu}{s_{\bar{x}}} < +t_{.05}$$

or that the value of 't' will lie within the limits shown. We can rearrange the above by multiplying through by $s_{\bar{x}}$, in which case we get: -

$$-t_{.05}s_{\bar{x}} < \bar{X} - \mu < +t_{.05}s_{\bar{x}}$$

By transposing and changing signs, we have: -

$$\bar{X} - t_{.05}s_{\bar{x}} < \mu < \bar{X} + t_{.05}s_{\bar{x}} \tag{6.4}$$

This is a familiar expression (see page 38) except that we now have substituted $t_{.05}$ for 2 in the case of the 5% probability and we can use $t_{.01}$ for 3 if we want the probability to be 1% or 99:1. All we need now to determine μ is to find a numerical value for 't'. The values for 't' for various probabilities and degrees of freedom are found in Table 6.1.

Table 6.1 - Distribution of 't' for various probabilities and degrees of freedom.

Degrees of freedom	Probability of a larger value of 't', sign ignored						
	0.5	0.4	0.2	0.1	0.05	0.025	0.01
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925
3	0.765	0.978	1.638	2.353	3.182	4.176	5.841
4	0.741	0.941	1.533	2.132	2.776	3.495	4.604
5	0.727	0.920	1.476	2.015	2.571	3.163	4.032
6	0.718	0.906	1.440	1.943	2.447	2.969	3.707
7	0.711	0.896	1.415	1.895	2.365	2.841	3.499
8	0.706	0.889	1.397	1.860	2.306	2.752	3.355
9	0.703	0.883	1.383	1.833	2.262	2.685	3.250
10	0.700	0.879	1.372	1.812	2.228	2.634	3.169
11	0.697	0.876	1.363	1.796	2.201	2.593	3.106
12	0.695	0.873	1.356	1.782	2.179	2.560	3.055
13	0.694	0.870	1.350	1.771	2.160	2.533	3.012
14	0.692	0.868	1.345	1.761	2.145	2.510	2.977
15	0.691	0.866	1.341	1.753	2.131	2.490	2.947
16	0.690	0.865	1.337	1.746	2.120	2.473	2.921
17	0.689	0.863	1.333	1.740	2.110	2.458	2.898
18	0.688	0.862	1.330	1.734	2.101	2.445	2.878
19	0.688	0.861	1.328	1.729	2.093	2.433	2.861
20	0.687	0.860	1.325	1.725	2.086	2.423	2.845
25	0.684	0.856	1.316	1.708	2.060	2.385	2.787
30	0.683	0.854	1.310	1.697	2.042	2.360	2.750
50	0.680	0.849	1.299	1.676	2.016	2.310	2.678
60	0.679	0.848	1.296	1.671	2.000	2.299	2.660
70	0.678	0.847	1.294	1.667	1.994	2.290	2.648
80	0.678	0.847	1.293	1.665	1.990	2.284	2.638
90	0.678	0.846	1.291	1.662	1.986	2.279	2.631
100	0.677	0.846	1.290	1.661	1.982	2.276	2.625
120	0.677	0.845	1.289	1.658	1.980	2.270	2.617
∞	0.6745	0.8416	1.2816	1.6448	1.9600	2.2414	2.5758

Table 6.1 is taken from Table IV of Fisher: "Statistical Methods for Research Workers", published by Oliver and Boyd Ltd., Edinburgh, and by permission of the author and publishers; from Maxine Merrington's "Table of Percentage Points of the t-Distribution", Biometrika, 32:300 (1942), and from Bernard Ostle's "Statistics in Research", Iowa State University Press (1954).

Across the top of Table 6.1 you will notice the statement "Probability of a larger value of 't', sign ignored." Since 't' has its own distribution which approaches the normal as sample size increases, it stands to reason that the value of 't' will depend upon (1) degrees of freedom and (2) the probability level desired.

Let us look at the 't' values for a sample of 10 items and for the 5% and 1% probability levels. We have 9 degrees of freedom (d.f. = 9), located in the left-hand column; read across until you reach the probability levels specified. The 't' values are 2.262 and 3.250 for the 5% and 1% probability levels respectively. These figures mean that in 5% of subsequent samples from the same population, the variability will be such that we would expect the calculated 't' values will be greater than 2.262 five times in 100 trials. The sign is ignored and 't' could easily have been written $|t|$, meaning the absolute value of 't'. Similarly, in 1% of the subsequent samples, the value of 't' will be expected to be greater than 3.250.

In the beginning of this chapter, we used some approximate probability levels associated with $2s_{\bar{x}}$ and $3s_{\bar{x}}$; they were 19:1 and 99:1, corresponding to the 5% and 1% levels. However,

the original did not take sample size into account because we had not been introduced to 't' at that point. You can now verify that the exact 5% and 1% levels of probability correspond to d.f. = 60 and 13 respectively. This is a far more refined method of determining fiducial intervals or limits and you will find the table of 't' to be increasingly useful as you progress through basic statistics.

Determining N for a given value of $s_{\bar{x}}$

An important use of standard error of the mean is in the determination of how many sampling units are needed to ensure that a sample mean is within a particular limit of accuracy. We use a sample mean to estimate the population mean, so it is important to know how closely we can estimate the population parameter. In most sampling problems, it is not necessary to estimate the population mean to a high degree of accuracy, but it is important to know within what limits it lies. The mean of the population is not a variable; it is a fixed, though unknown quantity. Remember that we are referring to an infinite population now, not a finite one in which it is possible to measure all the sampling units.

Forest inventory is a good example of a sampling process which uses a stated value of the standard error. Volumes calculated from sample plots constitute a sample of the population of sampling units. We might decide that a 5% sample (by area) is adequate, and proceed on that basis until the required number of plots has been secured. If the forest area is uniform as to site and age classes, a very good estimate of the population mean will result because the standard deviation (and hence the standard error) will be low in value. However, if there is much variation between stand types, the standard error will be high, giving a less accurate estimate of the population mean. How are we to know when enough sampling units have been measured? Sampling for forest inventory is an expensive process, and for the sake of economy, we do not want to sample excessively.

It is usual to specify that the sampling will be continued until the standard error of the mean is within ± 10% of the sample mean. Other limits may be imposed according to the sampling problem, but ± 10% is a very common figure. In addition to certain limits imposed upon the value of standard error, we also must specify the probability levels which must be met. The following example will illustrate the problem.

An area of 250 acres is sampled with 26 one-fifth acre plots. Volumes in board feet are calculated for each plot. A prerequisite of the survey is that the standard error of the mean ($s_{\bar{x}}$) shall be within ± 10% of the sample mean volume at the 5% level. How many plots must be taken to satisfy this requirement?

Solution

First, we must obtain the mean volume of the 26 plots. A total of 41,704 board feet was measured, giving a mean of: -

$$\bar{X} = \frac{41,704}{26}$$

$$= 1604 \text{ board feet}$$

Let us also assume that the standard deviation for the 26 plots is ± 620 board feet.

Standard error of the mean is therefore: -

$$s_{\bar{x}} = \frac{620}{\sqrt{26}}$$

$$= \frac{620}{5.1}$$

$$= \pm 121.5 \text{ board feet}$$

Obviously, the standard error of ± 121.5 board feet is within ± 10% of the sample

mean of 1604 board feet, but this does not answer the question. It would have satisfied the requirement of the standard error of the mean being within ± 10% of the sample mean 2 times out of 3, but we want to know if the requirement is satisfied at the 5% probability level which implies 95 times out of 100.

Let us set the standard error of the mean at 10% of the sample mean, and determine the number of plots required by rearranging the formula for $s_{\bar{x}}$ as follows: -

$$s_{\bar{x}} = \frac{s}{\sqrt{N}} \quad \text{or} \quad \sqrt{N} = \frac{s}{s_{\bar{x}}}$$

We need a 't' value, so add the necessary level of probability of 't' to the formula, which becomes: -

$$\sqrt{N} = \frac{t_{.05}s}{s_{\bar{x}}} \tag{6.5}$$

Substituting known values into the formula, we have: -

$$\sqrt{N} = \frac{2.060\ (620)}{160.4} \qquad \text{(see footnote)}$$

$$= \frac{1277.20}{160.4}$$

$$= 7.96$$

$$N = 63.3 \quad \text{or} \quad 64 \text{ plots}$$

From the calculations, we conclude that it would take 64 plots in the 250 acre tract to obtain the requirement that the standard error be with ± 10% of the sample mean, with a probability of 95 times out of 100 or at the probability level of 5%. Of course, we must assume that the additional 38 plots (64 - 26) do not vary in volume any more than the original 26 plots. After measuring the 64 plots, it would be wise to recalculate the mean, standard deviation and standard error to make sure that we are within the limits originally specified.

As an additional exercise, we might calculate the number of plots required to have the standard error within ± 10% of the sample mean at the 1% probability level. The calculations are: -

$$\sqrt{N} = \frac{t_{.01}s}{s_{\bar{x}}}$$

$$= \frac{2.787\ (620)}{160.4} \qquad \text{(see footnote)}$$

$$= \frac{1727.94}{160.4}$$

$$= 10.77$$

$$N = 116.19 \text{ or } 117 \text{ plots}$$

Tests of hypotheses

While it is of interest to determine fiducial limits by using 't', this statistic really comes into its own when we compare the means of two different samples. We want to know whether

---

The figure 2.060 is from the 't' table for d.f. = 25 and a probability of 0.05.

The figure 2.787 came from the 't' table for d.f. = 25 and probability of .01.

the numerical difference between two means is sufficient for us to say that the means came from two <u>different</u> populations, or whether they are means from the same population. Remember that in a normally-distributed population, we can have two successive means which differ numerically. See page 37 for the distribution of means. The further apart the means are, the less likelihood is there that they came from the same population. Once again, we can not make an arbitrary decision, but must follow a rigid set of rules and abide by the results. Look at Figure 6.2; from it, we have no way of telling whether the means of A and B represent samples from the same or different populations. They could be samples from the same population as in Figure 6.3a or from different populations as in Figure 6.3b.

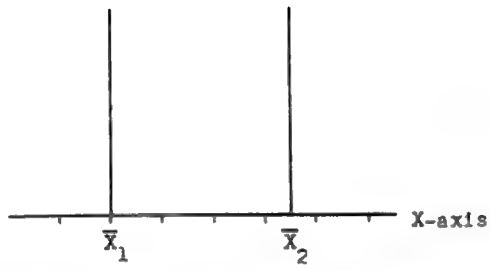
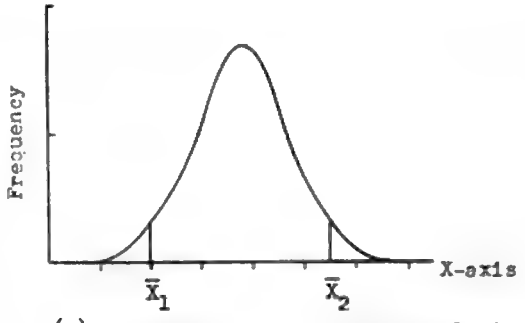


Figure 6.2 - Two sample means plotted on the X-axis.

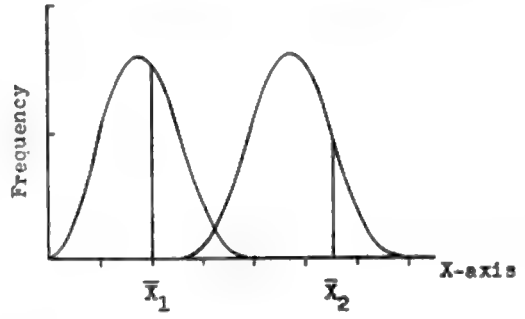
Statisticians are not gamblers and, in order that two means can be said to represent two different populations, there must not be more than 1 chance in 20 - the 5% probability level - that the difference is due to chance. Put another way, a <u>significant</u> difference exists between two means when a calculated 't' value exceeds the tabulated 't' value for the 5% probability level, when the proper degrees of freedom have been taken into account. Notice the word "significant" which was underlined; a proper understanding of this is important. A difference is said to be significant at the 5% level when there is not more than 1 chance in 20 that the difference is due to chance. A difference at the 5% level is denoted by *; if the difference were significant at the 1% level, it would be <u>highly significant</u> and denoted by **.

Null hypothesis theory

It is common practice in statistics to think of differences between means in terms of the null hypothesis theory. Let us assume that we are conducting an experiment in which we have used two strengths of a certain chemical to test their effect on height growth of pine seedlings. At the present time, it does not matter what the chemical is, but we must try to have all other factors influencing height growth the same for both groups of seedlings treated. More will be said about this in the chapter on Experimental Design. At the end of the experiment, each seedling is measured and a mean obtained for the seedlings treated by each strength of chemical. We now have $\bar{X}1$ and $\bar{X}2$, with the difference being shown as $\bar{X}1 - \bar{X}2$; we are not interested in whether $\bar{X}1$ is larger than $\bar{X}2$ or vice versa, so we could just as easily have shown the difference as $\bar{X}2 - \bar{X}1$ or $|\bar{X}1 - \bar{X}2|$, the absolute value of the difference.



(a) Two means from the same population



(b) The same two means but from different populations.

Figure 6.3 - Distribution of two means from the same or different populations.

At this point, we set up a hypothesis that there is no difference between the means of the populations from which the samples were drawn. In other words, the hypothesis is: -

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0 \qquad (6.6)$$

This being our hypothesis, we set out to determine whether we were right or not. If we can show no significant difference, we accept the null hypothesis; if not, we reject it and accept an alternate hypothesis of H_A: $\mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$. The evidence upon which we base our decision is by a comparison of 't' values - one calculated and one shown in the 't' table for the 5% probability level.

Types of errors

Since we are admitting the fact that the 5% probability level establishes a significant difference, we are also admitting that we could be wrong in our decision 5 times in 100. What if the experiment leads us to reject the null hypothesis when, in fact, it was true? We have committed what is called a Type I error. If the reverse is true, that we accept the null hypothesis when it is false, we commit a Type II error. The following tabulation will illustrate this point.

H_0	H_A	Decision on H_0	Decision on H_A	Correctness of decision	Probability should be
True	False	Accept	Reject	Right	High
True	False	Reject	Accept	Wrong, Type I error	Low
False	True	Accept	Reject	Wrong, Type II error	Low
False	True	Reject	Accept	Right	High

You remember that the statistic 't' had the formula: -

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

for single samples. Now we are dealing with two sample means so that the original formula for 't' will not work.

For the null hypothesis $\mu_1 - \mu_2 = 0$, we set up the following: -

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

and since $\mu_1 - \mu_2 = 0$ by our hypothesis, we have: -

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}_1 - \bar{x}_2}} \qquad (6.7)$$

The only portion of the formula with which you are not familiar is $s_{\bar{x}_1 - \bar{x}_2}$ which is called standard error of a difference. It can also be written as $s_{\bar{d}}$ where $\bar{d}$ is the difference between two means. Standard error of a difference is calculated in the same way as the regular standard error except that we use differences between means as the final criterion. The statistic $s_{\bar{d}}$ is an estimate of $\sigma_{\bar{d}}$.

An example of testing H_0: $\mu_1 - \mu_2 = 0$

Taking our original experiment of using two different strengths of chemicals on pine seedlings to determine whether there is a difference in height response, we might have data as shown in Table 6.2, which is a simulated experiment; the data are reduced and modified for easy computation. The table also shows the calculations necessary to determine $s_{\bar{x}_1 - \bar{x}_2}$.

In this experiment, we select a number of seedlings at random from each population; we have no reason to compare the height response of a particular seedling from those receiving the 10% solution with one receiving the 50% solution. The seedlings may have been grown from seed collected from different trees and an inherited characteristic of a genetic type could cause some change in height response. Methods are available to test differences between items which have been paired; pairing is indicated if we know that the individuals will react the same if treated the same. If we had two successive measurements on the same individual, we

would pair the measurements one against the other. For those interested in learning the techniques of determining differences between paired individuals, Snedecor (1956) has an excellent treatment of tbe subject.

With equal numbers of items in each group, the pooled variance is given by the formula: -

$$s^2 = \frac{\Sigma x^2}{2(n-1)} \tag{6.8}$$

where s^2 = pooled variance (variance is the name given to the square of standard deviation)

Σx^2 = sum of the pooled sum of squares

n = number of items in each group.

Table 6.2 - Data for computing the effect of two treatments on equal sized groups, items selected at random.

10% solution			50% solution		
Height response (X_1)	x_1	x_1^2	Height response (X_2)	x_2	x_2^2
mm	mm		mm	mm	
40	- 1	1	45	- 2	4
46	+ 5	25	52	+ 5	25
36	- 5	25	39	- 8	64
38	- 3	9	44	- 3	9
47	+ 6	36	54	+ 7	49
40	- 1	1	45	- 2	4
38	- 3	9	45	- 2	4
40	- 1	1	46	- 1	1
38	- 3	9	47	0	0
47	+ 6	36	53	+ 6	36
Sums: 410	0	152	470	0	196
Means: 41			47		

Treatment	n	d.f.	Sum	Mean	Sum of squares
50% solution	10	9	470	47	196
10% solution	10	9	410	41	152
Sum:	20	18	Diff. = 6		348

$$s^2 = \frac{\Sigma x^2}{2(n-1)}$$

$$= \frac{348}{18} = 19.3 \text{ mm.}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2(19.3)}{10}} \tag{6.9}$$

$$= \sqrt{3.86} = \pm 1.965 \text{ mm.}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

$$= \frac{47 - 41}{1.965}$$

$$= \frac{6}{1.965}$$

$$= 3.051 \text{ **}$$

The calculated value of 't' is 3.051 and the tabulated value for 2(n - 1) degrees of freedom is 2.878 at the 1% level. Our calculated value is larger than the tabulated value, so we conclude that there is a highly significant difference between the height response of pine seedlings as a result of the two strengths of chemicals used. The original hypothesis was, of course, that the difference between the two means is zero, or H_0: $\mu_1 - \mu_2 = 0$. On the basis of the experiment, we reject the null hypothesis and accept the alternative one that H_A: $\mu_1 - \mu_2 \neq 0$.

A convenient method of determining the value of 't' for two equal-sized groups is: -

$$t = \bar{X}_1 - \bar{X}_2 \sqrt{\frac{n(n-1)}{\Sigma x^2}} \qquad (6.10)$$

in which Σx^2 is the pooled sum of squares. Taking the data in the previous problem, we have: -

$$t = 47 - 41 \sqrt{\frac{10(9)}{348}}$$

$$= 6 \sqrt{\frac{90}{348}}$$

$$= 6 \sqrt{.2586}$$

$$= 6(0.5085) = 3.051 \text{ **}$$

as before.

Another formula which can be used when there are equal numbers of individuals in each groups is: -

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}} \qquad (6.11)$$

in which $s_{\bar{x}_1 - \bar{x}_2}$ = standard error of the difference between the two means

$s^2_{\bar{x}_1}$ = standard error of the first group squared

$s^2_{\bar{x}_2}$ = standard error of the second group squared

The standard error terms in Formula 6.11 come from the calculation of standard deviation as before. Let us re-examine the data presented in Table 6.2 in light of the formula just introduced.

Group 1 - 10% solution	Group 2 - 50% solution
$n_1 = 10$	$n_2 = 10$
$\Sigma x_1^2 = 152$	$\Sigma x_2^2 = 196$
$s_1^2 = \frac{152}{9}$	$s_2^2 = \frac{196}{9}$
$= 16.89$	$= 21.78$
$s_{\bar{x}_1}^2 = \frac{16.89}{10}$	$s_{\bar{x}_2}^2 = \frac{21.78}{10}$
$= 1.689$	$= 2.178$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{1.689 + 2.178}$$

$$= \sqrt{3.867}$$

$$= \pm 1.965$$

The answer is the same as we determined previously by Formulas 6.8 and 6.9.

You will notice that the standard deviation was obtained slightly differently than we have been used to; rather than use the square root sign, we omitted it and squared s. This is a term called 'variance' and is nothing more than the square of standard deviation. Similarly, in order to simplify the calculations, we used the square of standard error and extracted the square root last. You may find that this method has some advantages.

Analysis of two groups of unequal numbers

It is not possible at all times to have equal numbers of individuals in each group. An experiment may be conducted in which the effect of the number of hours of light on total weight of plants is desired. Each plant is put in a pot, say 15 of which were subjected to 8 hours of artificial light and 15 to 17 hours. As far as possible, all other conditions which affect growth are kept the same for each group. During the experiment, which may run for 10 weeks, some plants in each group die from unknown causes or may be destroyed accidentally. Can we use the same procedures for computing the significance of the differences between means as we did previously? Obviously not, because we no longer have equal numbers in each group.

If Group 1 has n_1 individuals, the degrees of freedom would be $n_1 - 1$; Group 2 having n_2 individuals, has $n_2 - 1$ degrees of freedom and the total degrees of freedom is $(n_1 + n_2 - 2)$. The pooled sum of squares is the sum of the individual sum of squares or: -

$$\frac{s^2}{n_1} + \frac{s^2}{n_2} = s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

$$= s^2 \left(\frac{n_1 + n_2}{n_1 n_2} \right)$$

and the standard error of the difference between the means is: -

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{n_1 + n_2}{n_1 n_2} \right)} \qquad (6.12)$$

The example detailed in the next section will clarify these points.

Example of two groups with unequal numbers in each group

Using the example of 15 plants subjected to 8 hours of artificial light and 15 plants to 17 hours, at the end of a 10-week period, the plants were removed from the pots, dried uniformly and then weighed. During the experiment, 3 plants from Group 1 and 1 plant from Group 2 died from unknown causes, leaving $n_1 = 12$ and $n_2 = 14$. The corresponding d.f. are 11 and 13 respectively. The data and computation are included in Table 6.3.

Table 6.3 - Data for computing the effect of hours of artificial light on plant weight for unequal-sized groups.

	8 hours light			17 hours light		
	Dry weight X_1	x_1	x_1^2	Dry weight X_2	x_2	x_2^2
	13	- 2	4	15	- 3	9
	15	0	0	19	1	1
	15	0	0	18	0	0
	10	- 5	25	16	- 2	4
	17	2	4	17	- 1	1
	16	1	1	22	4	16
	14	- 1	1	19	1	1
	18	3	9	21	3	9
	10	- 5	25	18	0	0
	16	1	1	16	- 2	4
	16	1	1	20	2	4
	20	5	25	22	4	16
				15	- 3	9
				14	- 4	16
Sums:	180	0	96	252	0	90
Means:	15			18		

Source	n	d.f.	Sum	Mean	Sum of squares
17 hours of light	14	13	252	18	90
8 hours of light	12	11	180	15	96
Sum:		24	Diff. = 3	Σx^2	186

$$\text{pooled } s^2 = \frac{186}{24} = 7.75$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \left(\frac{n_1 + n_2}{n_1 \, n_2} \right)}$$

$$= \sqrt{7.75 \left(\frac{26}{168} \right)}$$

$$= \sqrt{1.199}$$

$$= \pm 1.095$$

$$t = \frac{18 - 15}{1.095}$$

$$= \frac{3}{1.095}$$

$$= 2.739*$$

The calculated 't' is significant at the 5% level (but not at the 1% level) and we would fail to accept the null hypothesis and conclude that the two treatments were significantly different.

The value of 't' for this experiment having different numbers of individuals in each group, can also be obtained from: -

$$t = \bar{X}_1 - \bar{X}_2 \sqrt{\frac{n_1 \, n_2 (n_1 + n_2 - 2)}{(n_1 + n_2)\Sigma x^2}} \qquad (6.13)$$

$$= 3 \sqrt{\frac{(168)(24)}{(26)\ (186)}}$$

$$= 3 \sqrt{\frac{4032}{4036}}$$

$$= 3 \sqrt{\ 0.8337}$$

$$= 3\ (0.9130)$$

$$= 2.739^{*} \text{ as before}$$

Do not use the alternate formula for standard error of a difference between two means (Formula 6.11) when comparing two groups with unequal means. The results will not be the same as if you use Formula 6.12.

Chapter 7

SAMPLING TECHNIQUES

Before discussing various sampling techniques and their applications to forestry, it might be well to review the terms and definitions. A schematic diagram is given in Figure 7.1 to show the relationship which exists between sample statistics and population parameters.

Sampling is concerned with the collection of quantitative or qualitative information from sampling units which possess the characteristics in which we are interested.

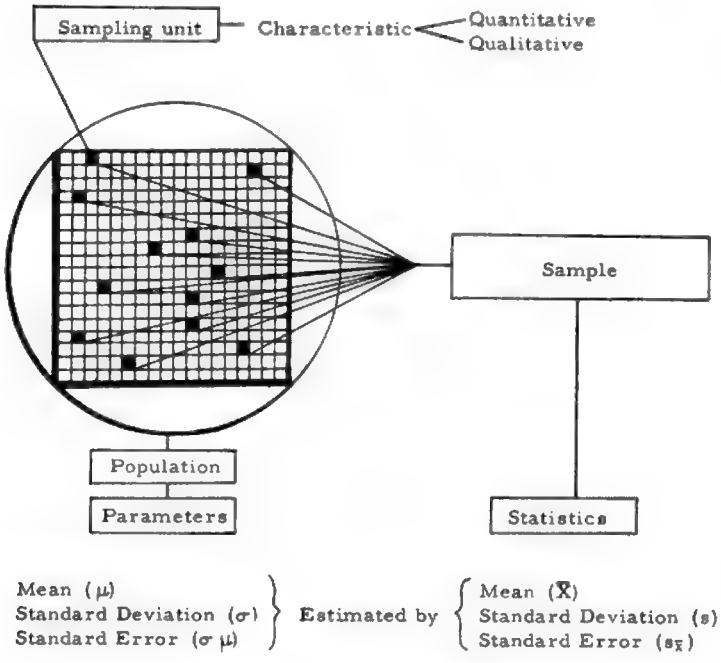


Figure 7.1 - Schematic representation of the relation between a sample and the population from which it was taken.

A quantitative characteristic is one which can be measured in terms of regular units of measure such as feet, inches, pounds, grams, or enumerated as to the number of occurrences.

A qualitative characteristic can be described but not measured in the ordinary sense. Examples are crown class, type of defect, soil texture, soil erosion, topography etc. There are methods of transforming qualitative characteristics into numerical form so they can be used in statistical analyses. Coile (1952), Zahner (1957) and others have dealt with this problem in site investigations. Nash (1963) used a rating system to transform the four qualitative variables of (1) slope and aspect (2) soil texture (3) slope position and (4) soil consistence into a numerical rating for use in a multiple regression to predict site index of shortleaf pine in Missouri.

A population is the sum total of all the sampling units within the given definition of the population. In normal statistical theory, the population is considered to be infinite in size. In practice, we often deal with finite populations containing only a certain number of sampling units. Certain corrections have been applied to statistical theory when dealing with finite as opposed to infinite populations, but in this beginning treatment of the subject, we will consider populations as being infinite.

The sample consists of a number of individual sampling units which not only possess the desired characteristic but which, taken together, are representative of the population. You will remember that there are three requisites for a sample: -

1. it must be obtained without bias.
2. it must be representative of the population.
3. it must be of sufficient size to ensure that the variation in the population is reflected in the sample.

Since the application of statistical theory, particularly the estimate of sampling error, demands that the sample be selected without bias, we will consider random sampling methods first.

Random sampling

In a random sample, each sampling unit in the population must have an equal opportunity of being selected. This is a basic principle. There are two general types of random sampling: -

1. replacement methods.

2. non-replacement methods.

The first implies that, once a sampling unit has been selected, it is replaced in the population and has another chance of being selected. In forestry applications, replacement sampling is not very important; it is unlikely that we would want a plot to be measured twice in a forest inventory or to measure the height of a seedling more than once just because its number came up again.

This being the case, we turn to non-replacement methods to obtain most samples in forestry. If a population consists of N sampling units in a finite population, and a sample of n units is desired, we could either -

1. write the quantitative characteristic of each sampling unit (N in all) on a slip of paper. The slips of paper can then be thoroughly mixed and one slip drawn out at a time until n slips have been drawn.

2. assign consecutive numbers to the N sampling units and prepare slips of paper numbered from 1 to N. Then draw out n slips and record the quantitative characteristic of each by matching the numbers against the original list.

These two methods are tedious and time-consuming; there is also the danger that the probability of the draw will not be consistent because of varying weights and thicknesses of the sampling media. These methods are satisfactory for the local bazaar or raffle, but for a completely random sample, we need greater refinement in technique.

Use of table of random numbers

By far the most direct way of securing a random sample is by using a table of random numbers. Table A.1 in the Appendix contains 2000 numbers, ranging from 0 to 9, which have been listed in a completely random order. Most texts on statistical methods contain such a table. The numbers are arranged in groups of five's horizontally and vertically for convenience only.

We shall use a small excerpt from Table A.1 to illustrate the method of selecting numbers for a random sample. The random numbers in Table 7.1 were copied from the table by starting at Column 00 Row 10.

Table 7.1 - Random numbers taken from Table A.1 in the Appendix

96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
47534	09243	67879	00544	23410	12740	02540	54440	32949	13491

(a) to select ten 2-digit numbers

Enter the table at any point and read the numbers in pairs from right to left, left to right, up or down. It makes no difference in which direction we proceed, as long as we are consistent for the sample. For the sake of argument, let us start at the sixth figure from the left on the second row down in Table 7.1 above. The first ten pairs of numbers will constitute our sample. They are: -

88, 04, 05, 33, 64, 71, 72, 64, 56 and 90.

Selecting groups of two's from the table of random numbers is satisfactory for a sample where $N < 100$, or where the numbers refer to columns and rows in a grid system in which there are less than 100 columns or rows.

(b) to select ten 3-digit numbers

A sample of ten items consisting of 3-digit numbers follows the same pattern as for 2-digit numbers. If the original sampling units are numbered from 1 to 999, it is easy to select a sample of ten by taking the numbers from the table in groups of three's. Thus, starting at Column 2 Row 3 and reading down, we have 627. Now starting at the top of Column 3 and reading down again, we have the following sets of three's: -

627 (first group), 733, 155, 511, 347, 814, 183, 507, 872 and 960.

(c) to select a sample of ten 3-digit numbers restricted to a specific upper limit.

At times, we have a problem of sampling which we require ten sampling units to be taken from a population having a total of 500; this is a finite population having a specific number of sampling units. To select a sample of ten from this population will require groups of three's from the table of random numbers with the restriction that any group of three numbers which exceeds 500 can not be accepted. If one set is discarded, we proceed to the next set until the number fits our requirements. Starting at Column 11 Row 1 and reading from right to left, we have: -

556, 594, 410, 547, 361, 348, 338, 667, 923, 930,

532, 492, 708, 334, 357, 880, 405, 336 and 471.

The acceptable numbers are underlined.

(d) to select a random sample of 35 from an irregularly-shaped forest area.

Assuming that a random sample of plots is required from an irregularly-shaped forest area, we are faced with a problem of sampling particular points within the forest. This is a common problem in some phases of forest inventory; the solution is simple. Grid the area into conveniently-sized squares so that it encompasses the whole area. Then select from the table of random numbers those groups of two's which will fit the area by the intersection of column and row numbers, discarding those which fall outside the forested area or which are beyond the limits of the grid system.

Figure 7.2 is the result of such a sampling system on an irregularly-shaped area, with the sampling points shown. The question of whether this is the most efficient method of sampling is not the question at the moment. Where a completely random sample is required, this is a good way of obtaining it.

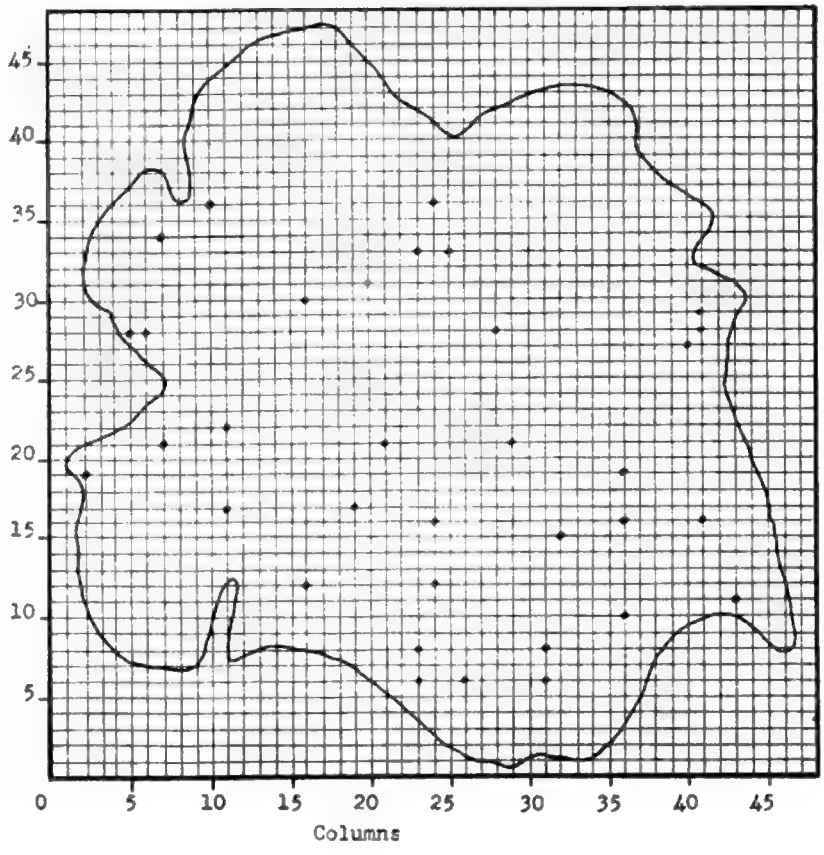


Figure 7.2 - A random sample of 35 taken from an irregularly-shaped area.

Example of a forest inventory problem

Forest inventories are taken for the express purpose of determining volume of standing trees, or other quantitative data, in order that the forester may make meaningful decisions regarding the management of the resource. Volumes may be measured in many units- total cubic feet, board feet, cords, merchantable cubic feet; diameter distributions may be determined from a sampling process and the results applied to the whole forest; defect studies may be conducted to determine the extent of merchantable volume loss. The purposes of conducting a forest inventory are almost innumerable; the most important consideration is that the sampling be done according to acceptable statistical procedures and that the sampling errors reflect the variability of the characteristic being sampled. This implies that non-sampling errors such as errors in measurement, in calculation or others are eliminated so do not enter into the figure for sampling error. Control procedures initiated prior to and during the inventory will ensure that human errors are not in the calculations.

As volume per acre is usually a final result of a forest inventory, we can use it as an example and can approach the problem in one of two ways:

1. we could perform a 100% enumeration of all trees in an area, compute volumes on a per acre basis and then sample the forest by a random selection and check our estimates against the 100% cruise data. A hypothetical 400-acre forest, with volume for every acre is shown in Figure 7.3. This will be used for both random and stratified random sampling

						C	O	L	U	M	N									ROW
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
8	7	12	9	6	8	12	14	7	4	5	4	3	7	4	2	4	6	7	7	1
16	18	9	10	7	9	14	10	6	3	4	8	5	6	5	3	3	4	6	6	2
22	15	17	14	12	13	16	9	4	5	6	5	7	7	6	5	4	5	7	9	3
19	16	13	10	8	9	12	9	8	7	9	10	9	8	3	3	7	10	13	10	4
16	12	12	9	5	7	10	12	6	8	7	6	6	5	6	4	6	7	9	10	5
19	16	11	7	8	10	14	10	16	15	17	16	18	17	19	22	24	20	19	15	6
22	18	16	19	12	13	17	12	18	21	20	19	25	22	24	27	20	17	22	14	7
25	14	16	12	13	15	16	14	23	25	28	33	30	27	31	25	22	20	24	29	8
20	17	14	16	17	19	20	24	25	27	32	40	42	35	35	37	29	28	27	35	9
25	22	19	25	29	32	36	30	35	40	41	47	45	40	32	33	27	27	31	30	10
29	30	27	32	37	42	45	45	43	51	53	52	48	39	42	37	33	29	27	20	11
31	35	38	45	47	44	49	51	50	58	56	50	41	43	30	27	29	23	21	19	12
28	33	30	37	40	39	42	46	51	48	44	40	37	30	35	22	21	17	14	16	13
26	29	32	35	35	31	29	32	40	37	35	31	16	19	20	11	12	10	13	12	14
22	28	31	29	28	27	25	27	16	22	19	16	8	4	2	3	6	5	7	10	15
21	22	28	26	31	27	22	24	20	14	10	8	6	6	4	7	4	3	1	5	16
17	26	30	31	31	27	25	33	12	8	7	6	6	4	2	1	0	0	0	0	17
17	24	31	33	27	22	19	17	14	10	9	7	3	0	3	2	0	4	2	0	18
19	19	21	17	15	16	16	19	16	13	10	12	8	3	1	0	3	2	5	3	19
23	21	17	14	13	19	23	17	12	11	6	9	7	3	1	0	1	1	2	4	20

Figure 7.3 A hypothetical 400-acre forest showing volume in hundreds of board feet for each acre on the block.

2. we could assume nothing regarding the total volume for the area but establish a sample and prepare volume estimates.

The latter approach is more practical. A random sample, if correctly carried out, will provide an unbiased estimate of volume for the entire area and will include such statistics as the mean, standard deviation, standard error of the mean and confidence limits according to a pre-determined probability. The sampling statistics are the sole basis for our estimate and very seldom is a sample compared to the results of a 100% enumeration.

The question is often asked: When should one sample and when should one do a 100% tally? The answer will depend on a number of considerations such as the characteristic being sampled, size of the area, time and money available and so on. For most forest inventory problems involving area, a general rule is that if the area is larger than 40 acres, use a sampling procedure. For smaller areas, it might be almost as fast to measure all trees in well-defined strips.

Returning to the problem in forest inventory sampling, let us assume that we have the following specifications:

1. size of area -	400 acres in a square block
2. size of sample -	10% or 40 acres to be measured in 1-acre sampling units
3. type of sample -	random
4. standards of accuracy -	standard error of the mean to be within + 10% of the sample mean with a probability of 2 out 3 (P = 0.33)

We can start by laying out grid lines in the field so that we have 20 rows and columns with each square occupying one acre. By using a table of random numbers, we can select a sample of n = 40 and then, by measuring the trees in these 1-acre plots, arrive at a volume for each acre sampled. A schematic representation of the area and the volumes in the 1-acre plots is shown in Figure 7.4. The original 40 plots are those with the light borders and those with the heavy border are additional plots required under Specification 4 above and will be explained later.

At this point, it is well to mention that the problem deals with a finite population. Most forest inventory problems are concerned with this type rather than with an infinite population. When dealing with an infinite population, the divisor for the standard error of the mean formula is $\sqrt{n}$:

where $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

and $s_{\bar{x}}$ is the standard error of the mean

s is the standard deviation

n is the number of items in the sample.

However, for a population which is limited in size, having a total of N sampling units, from which a sample of n is drawn, an estimate of the standard error of the mean is given by:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

in which $\left(1 - \frac{n}{N}\right)$ is the finite population correction factor.

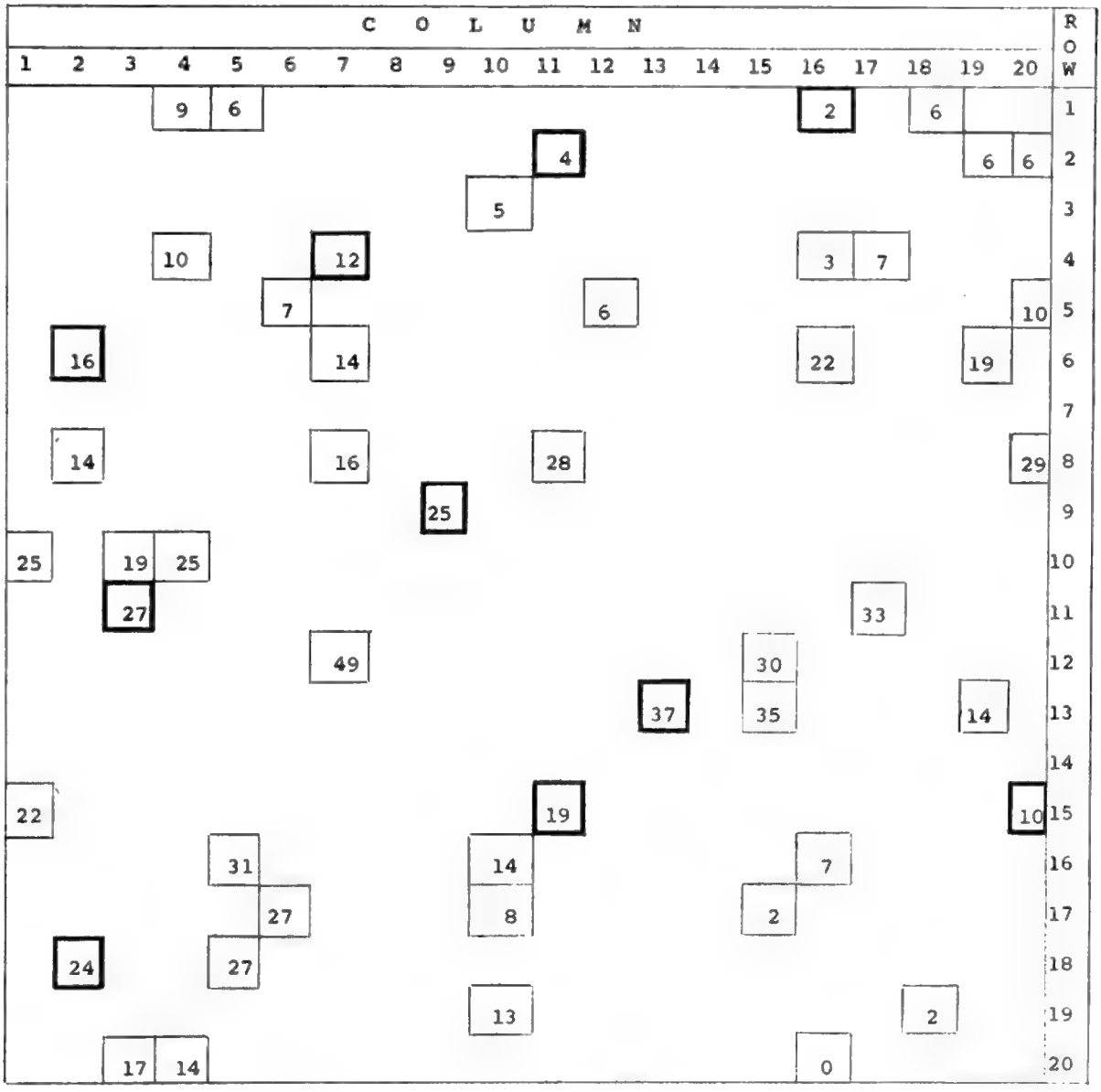


Figure 7.4 - Volume in hundreds of board feet measured at 40 sampling points (light border) and 10 additional points (heavy border), all selected at random.

As sample size increases, the standard error of the mean, $s_{\bar{x}}$, will decrease and the estimation of population parameters becomes more accurate. The mean of a 100% enumeration has no sampling error because n=N and the quantity $\left(1 - \frac{n}{N}\right)$ becomes zero.

Returning to Figure 7.4, we can determine the sample statistics for the 40 plots. They are:

		Board feet 00's
Mean volume ($\bar{V}$)		15.92
Standard deviation (s)	$\pm$	11.16
Standard error of the mean ($s_{\bar{V}}$)	$\pm$	1.61

The standard error of the mean volume ($s_{\bar{v}}$), does not comply with Specification 4; the standard error of the mean was to be within + 10% of the sample mean volume. To fulfill this requirement, the standard error of the mean should be + 1.59 or less. We therefore must calculate how many additional samples are necessary. We use the formula from page 42:

$$\sqrt{n} = \frac{t \cdot s}{s_{\bar{v}}}$$

or, by removing the square root sign:

$$n = \frac{t^2 \ s^2}{s_{\bar{v}}^2}$$

where n = number of sampling units

t = value of t for a designated probability level (t = 1.0 for d.f. 39 and P = 0.33)

s = standard deviation of the original sample plots

$s_{\bar{v}}$ = standard error of the mean when $s_{\bar{v}}$ = 10% of $\bar{V}$

Substituting known values in the formula, we have:

$$n = \frac{(1.0^2) \ (11.16^2)}{1.59^2}$$

$$= \frac{124.55}{2.53}$$

= 49.22 or 50 sampling units

We need to measure an additional 10 plots to bring the standard error of the mean to + 10% of the sample mean. These 10 additional plots, selected at random, are shown with heavy borders in Figure 7.4. The revised statisitcs for the 50 plots are:

n = 50

$\bar{V}$ = 16.26

s = ± 11.09

$s_{\bar{v}}$ = ± 1.37 (after using the finite population correction factor)

Our standard error is now within the required limits.

A note of caution at this point. The additional plots may not reduce the standard error of the mean to the set standard. A random sample of 10 additional plots may cause a selection of sampling units with a much higher variation in volume than in the original sample. This can very well happen in a population which shows wide variation between extreme values. This problem usually is alleviated by the use of stratified random sampling (see next section).

Confidence limits for a probability of 2 out of 3 (P = .33) are:

$$\bar{V} \pm t_{.33} \ s_{\bar{v}} \quad \text{where } t = 1.00 \text{ for d.f.} = 49$$

$$16.26 - 1.00\ (1.37) < \mu < 16.26 + 1.00\ (1.37)$$

$$14.89 < \mu < 17.63$$

and the volume estimate for the 400-acre block will be:

$$14.89\,(400) < \mu < 17.63\,(400)$$

$$5956 < \mu < 7052$$

This estimate is the figure which the forest manager wants to know. The estimate will be correct with a probability of 2 times out of 3. If this probability is too low, we can determine the estimates for a probability of 95 times out of 100 (P = .05) and for 99 times out of 100 (P = .01). The data for these are:

a. for a probability of 95 times out of 100

$$\bar{V} \pm t_{.05}\, s_{\bar{v}} \qquad \text{where } t = 2.06 \text{ for d.f.} = 49$$

$$16.26 - 2.06\,(1.37) < \mu < 16.26 + 2.06\,(1.37)$$

$$16.26 - 2.82 < \mu < 16.26 + 2.82$$

$$13.44 < \mu < 19.08$$

and the volume estimate for the 400-acre block is:

$$13.44\,(400) < \mu < 19.08\,(400)$$

$$5376 < \mu < 7632$$

b. for a probability of 99 times out of 100

$$\bar{V} \pm t_{.01}\, s_{\bar{v}} \qquad \text{where } t = 2.70 \text{ for d.f.} = 49$$

$$16.16 - 2.70\,(1.37) < \mu < 16.26 + 2.70\,(1.37)$$

$$16.26 - 3.70 < \mu < 16.26 + 3.70$$

$$12.56 < \mu < 19.96$$

and the volume estimate for the 400-acre block is:

$$12.56\,(400) < \mu < 19.96\,(400)$$

$$5024 < \mu < 7984$$

The presentation of the results of the 100% enumeration on the 400-acre block has been delayed to this time so you could appreciate the sampling process on its own merits. The 100% enumeration resulted in a mean vol- of 18.57 hundreds of board feet and a total volume of 7429 hundreds of board feet.

Looking back at our sample of 50 acres, we find that the probability level of 2:3 (P = .33) did not include the true mean of the population but that the other two (P = .05 and P = .01) did. Unless we actually perform a 100% enumeration, we have no way of knowing whether the sample statistics include the true mean of the population at the various probability levels. Therefore, we use the statistics of a single sample to estimate the population parameters, knowing that our estimates are only as good as the sample dictates.

This example of random sampling does not take into account the fact that most forested areas are stratified by species, species groups, by volume, site or some other factors which affect growth. Pure random sampling is not a very efficient procedure in the field, even though it is statistically correct. A logical outgrowth of random sampling which deliberately reduces the standard error of the mean is stratified random sampling which is considered in the following section.

Stratified random sampling

To increase efficiency in sampling, plots may be selected so that the variation between plots is reduced by taking advantage of natural stratification. Examine Figure 7.5; this diagram represents a 100% cruise on a 400-acre tract with lines drawn around areas having volumes which show a lower range within blocks than for the whole tract. Any random sample within a stratum or block will have a lower standard deviation and standard error than for the whole population. The range of volume is: -

	Volume/acre		
Block	High	Low	Range
A	36	7	29
B	16	2	14
C	42	12	30
D	58	27	31
E	16	0	16
Entire tract	58	0	58

The process of stratification is a deliberate attempt to reduce sampling errors by restricting the range within the sub-populations.

Stratification can be achieved in a number of ways. Perhaps the most effective is by the use of aerial photographs which can be used to delineate stands of similar characteristics such as species, height, crown density and crown diameter. Each stand or group of stands with the same characteristics is considered a stratum for purposes of sampling.

						C	O	L	U	M	N									ROW
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
8	7	12	9	6	8	12	14	7	4	5	4	3	7	4	2	4	6	7	7	1
16	18	9	10	7	9	14	10	6	3	4	8	5	6	5	3	3	4	6	6	2
22	15	17	14	12	13	16	9	4	5	6	5	7	7	6	5	4	5	7	9	3
19	16	13	10	8	9	12	9	8	7	9	10	9	8	3	3	7	10	13	10	4
16	12	12	9	5	7	10	12	6	8	7	6	6	5	6	4	6	7	9	10	5
19	16	11	7	8	10	14	10	16	15	17	16	18	17	19	22	24	20	19	15	6
22	18	16	19	12	13	17	12	18	21	20	19	25	22	24	27	20	17	22	14	7
25	14	16	12	13	15	16	14	23	25	28	33	30	27	31	25	22	20	24	29	8
20	17	14	16	17	19	20	24	25	27	32	40	42	35	35	37	29	28	27	35	9
25	22	19	25	29	32	36	30	35	40	41	47	45	40	32	33	27	27	31	30	10
29	30	27	32	37	42	45	45	43	51	53	52	48	39	42	37	33	29	27	20	11
31	35	38	45	47	44	49	51	50	58	56	50	41	43	30	27	29	23	21	19	12
28	33	30	37	40	39	42	46	51	48	44	40	37	30	35	22	21	17	14	16	13
26	29	32	35	35	31	29	32	40	37	35	31	16	19	20	11	12	10	13	12	14
22	28	31	29	28	27	25	27	16	22	19	16	8	4	2	3	6	5	7	10	15
21	22	28	26	31	27	22	24	20	14	10	8	6	6	4	7	4	3	1	5	16
17	26	30	31	31	27	25	33	12	8	7	6	6	4	2	1	0	0	0	0	17
17	24	31	33	27	22	19	17	14	10	9	7	3	0	3	2	0	4	2	0	18
19	19	21	17	15	16	16	19	16	13	10	12	8	3	1	0	3	2	5	3	19
23	21	17	14	13	19	23	17	12	11	6	9	7	3	1	0	1	1	2	4	20

Figure 7.5 Stratification of volumes shown in Figure 7.3

If each stratified area were of the same size, we could establish the same number of plots in each and arrive at a 10% sample for each stratum and for the whole tract. Things do not usually happen so neatly in practice. We find that one classification may be much larger than another, and measuring the same number of plots in each block would give us unequal sampling percentages. The distribution of area by blocks in Figure 7.5 is as follows: -

Block	Area acres
A	91
B	80
C	101
D	57
E	71
	400

For purposes of sampling, these can be rounded off to 90, 80, 100, 60 and 70 so that for a straight 10% sample, we should measure 9 plots in Block A, 8 in Block B, 10 in Block C, 6 in Block D and 7 in Block E. This method ensures that the sample will be based on an area basis. However, this method ignores an important consideration in forestry; it is that areas having greater volume per acre (or value) should be sampled more heavily than those with less volume or value. The method of sampling to take volume into account is called the volume-area proportionate method. Taking each block and multiplying the area by its average volume per acre gives us the figures in Table 7.2.

Table 7.2 - Volume-area combinations of blocks from Figure 7.5.

Block	Average volume/acre fbm in 00's	Area acres	Volume x area
A	18	90	1620
B	7	80	560
C	25	100	2500
D	41	60	2460
E	5	70	350

Taking the lowest value in the last column and dividing it into the others gives us ratios of approximately 5 : 2 : 8 : 8 : 1 in order from A to E. Our 10% sample for the whole tract requires 40 sampling units or plots, so each of the ratios should be multiplied by a factor derived from the total number of plots required divided by the sum of the ratios or 40 ÷ 24 = 1.67. The number of plots to be measured in each block is therefore: -

Block	Factor	Number of plots
A	5(1.67)	8.35
B	2(1.67)	3.34
C	8(1.67)	13.36
D	8(1.67)	13.36
E	1(1.67)	1.67
		40.08

Since we can not take a fraction of a plot, the number of plots can be rounded off. The percent sample for each block can be determined as shown in Table 7.3.

Table 7.3 - Percent sample by blocks and for the whole tract.

Block	Number of plots	Area acres	Percent sample
A	8	90	8.88
B	3	80	3.75
C	14	100	14.00
D	13	60	21.70
E	2	70	2.86
Total	40		51.19
Mean			10.24

You will notice that the block with the highest average volume is sampled more heavily than the others. As a test of sampling errors by blocks, the number of plots indicated in Table 7.3 was selected at random within each block. The mean, standard deviation and standard error were calculated and then compared with the parameters for the whole tract. The results are shown in Table 7.4.

Table 7.4 Statistics for samples drawn at random from stratified blocks

Block	Number of plots	Volume in 00's fbm V	s	$s_{\bar{V}}$	Coefficient of variation %
A	8	16.87	± 3.40	± 1.10	20.1
B	3	5.00	± 0.81	± 0.44	16.5
C	14	25.14	± 5.04	± 1.15	20.0
D	13	44.00	± 7.82	± 1.67	17.8
E	2	3.00	± 0.81	± 0.55	27.0
400-acre block	40	16.26	±11.16	± 1.37	68.7

The appropriate finite population correction factors have been applied to the standard error values ($s_{\bar{V}}$) which were calculated with a probability of 2 out of 3 (P = 0.33).

The stratified random samples can be compared to the 40-acre sample taken from the entire block. It is particularly interesting to compare the coefficients of variation.

Systematic sampling

Although pure random sampling is the basis for statistical theory and estimates of sampling error, it is not an efficient method for sampling forested areas. Too much time is spent locating the sampling points in the field and travel time is non-productive. The establishment of field plots is an expensive proposition, so foresters turned to systematic sampling as a substitute for random sampling. It must not be thought that random sampling methods have been completely eliminated by the forestry profession; there are occasions and circumstances which warrant it. For instance, research projects which demand an accurate estimate of sampling errors require a random method of sampling.

Systematic sampling, in which the location of the sampling units is chosen by some predetermined pattern, has the characteristic of being unable to provide a precise estimate of standard error. If it is calculated, it is a measure of the maximum standard error rather than the average which is obtained by random sampling methods. Because of this, systematic sampling gives conservative results; from a practical point of view, this has some merit.

Consumer surveys are often carried out by systematic sampling from a telephone book, city directory or tax assessment rolls. TV program ratings are often established by a telephone questionnaire directed at persons selected systematically from a telephone book. This method automatically excludes those persons who own a TV set but do not have a telephone.

An element of randomness can be introduced into a systematic sampling method which will ensure that each sampling unit has an equal chance of being selected. There are two possibilities depending on the manner in which the original data are presented.

The first is concerned with a straight listing of items which can be numbered consecutively. Let us assume that we have 650 items listed and that we wish to obtain a sample by selecting every 10th item from the list. Arbitrarily deciding on items numbered 10, 20, 30 etc. has no element of randomness in it because it excludes the intervening numbers. However, if we were to pick a number from 0 to 9 at random, we would then be giving each number in the first 10 an equal opportunity of being selected. We can use the table of random numbers to select the initial sampling unit. If the number selected is 3, then the order of selection from the list would be 3, 13, 23, 33 ..... 643.

The second situation is concerned with the selection of sampling units from rows and columns in a block. Here we need to know, in advance, how many plots are to be selected from each column and how many rows will be utilized for the sample. As an example, we might determine that a plot is to be established every six chains on lines 10 chains apart. To fix a starting point, we can obtain a random intersection of a number from 1 to 6 to indicate the row and one from 1 to 10 to indicate the column. Select at random a figure from the table of random numbers; say it is 4. The sequence of plots within lines is therefore 4, 10, 16, 22 etc. chains from the border of the area. Now obtain a number from 1 to 10, for instance 8. The sequence of lines is therefore 8, 18, 28, 38 etc. The pattern of the sampling is represented in Figure 7.6.

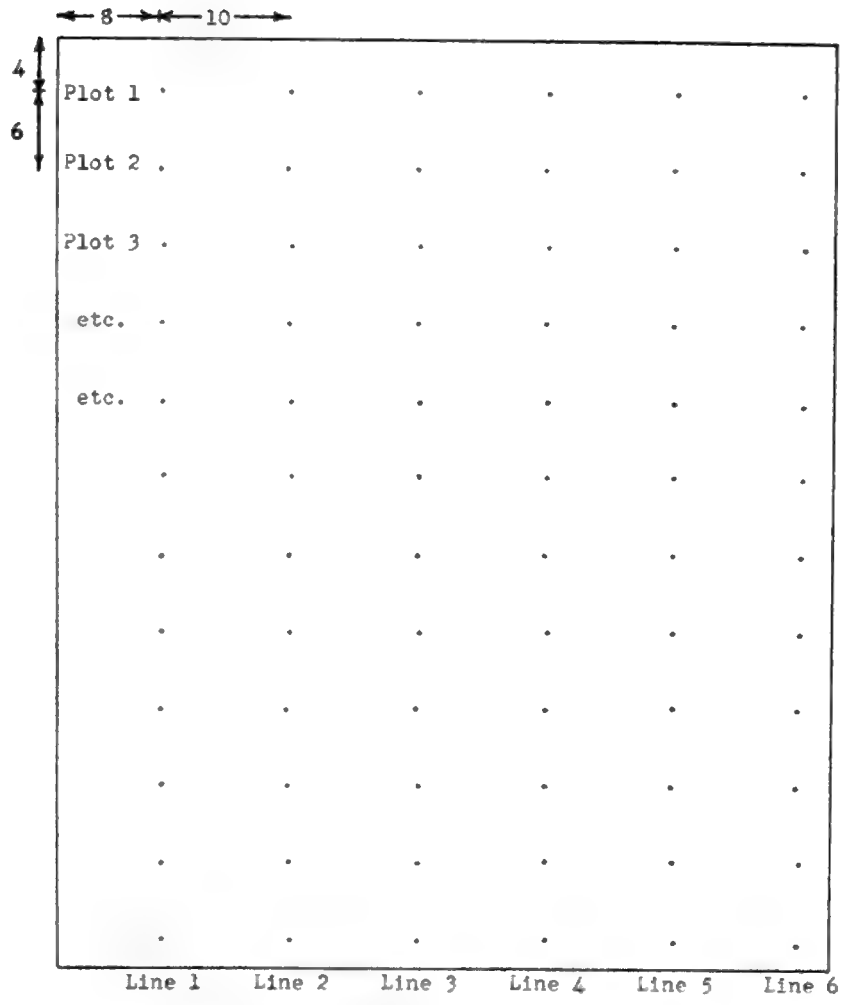


Figure 7.6 - Location of plots in a systematic sample spaced 6 chains by 10 chains.

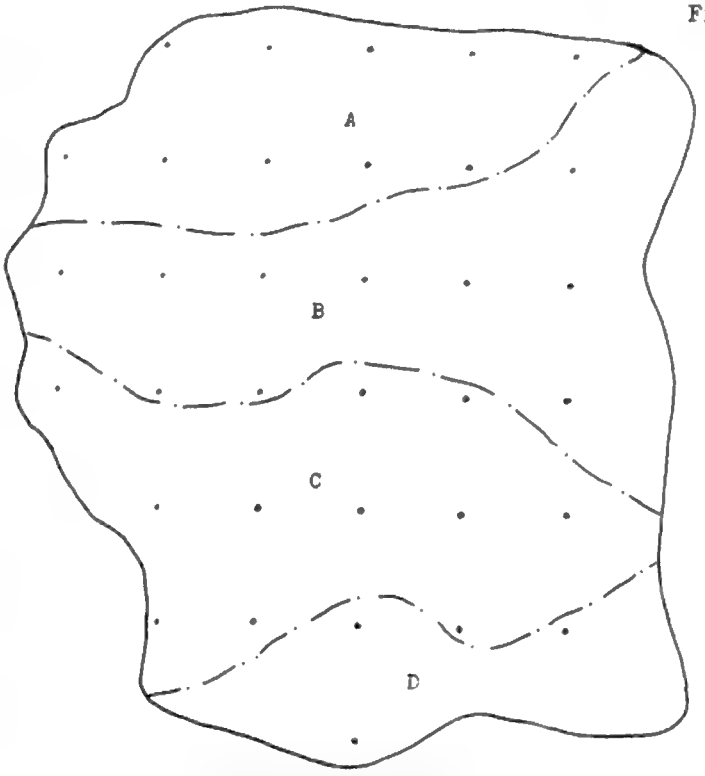


Figure 7.7 - Systematic sampling on a tract subdivided into strata.

Systematic sampling of stratified blocks requires considerable attention to block boundaries and is not as easy to carry out as for square or rectangular blocks. One possible solution is to design a systematic sample without reference to stratification lines and then to assign plots as they are measured to a particular stratum. This also ensures that the sampling will be done on an area basis with the larger strata being sampled more heavily than the smaller ones. Figure 7.7 is an example of systematic sampling for a large block containing four strata. The number and location of the plots fit the whole tract.

More sophisticated forms of sampling

We have covered unrestricted random sampling, stratified random sampling and systematic sampling. These are the most common types of sampling designs in forestry and are sufficient for the beginner to understand.

More sophisticated sampling designs are used, particularly in forest inventory programs. Difficulty of travel will sometimes dictate a modification of the unrestricted random sampling method. Cost of measuring a particular variable, when compared to a related variable may preclude unrestricted random sampling.

Other sampling designs, commonly used in forest inventory are:

1. cluster sampling
2. multiphase sampling
3. multistage sampling

Cluster sampling will be presented in some detail as it is an efficient sampling design, particularly in areas where access is difficult.

Cluster sampling

Cluster sampling was designed to accomodate the situation where transportation facilities - roads, rivers, trails - are very few such as in mountainous areas or in tropical regions where roads may not exist. A field crew may take from one to four days or more just to reach a designated sampling point in the forest. With a little extra time, more than one sample can be taken in the vicinity.

Let us take an example where there are n clusters, each having m sub-plots.

An estimate of the population mean may be derived by obtaining the average of the measured values on all sampled sub-plots of all sampled clusters. It is assumed that the sub-plots in each cluster are selected at random from the total possible number of sub-plots in a cluster and that the clusters are chosen at random from the total possible number of clusters in the population.

1. the mean for each cluster is obtained from:

$$\bar{X}_c = \frac{\bar{X}_{s1} + \bar{X}_{s2} + \ldots + \bar{X}_{sm}}{m}$$

where $\bar{X}_c$ is a cluster mean

$\bar{X}_{s1}$, $\bar{X}_{s2}$ etc. are the mean values within each of the sub-plots

m is the number of sub-plots sampled in each cluster

2. the estimate of the grand cluster mean provides an estimate of the population mean by:

$$\bar{\bar{X}} = \frac{\bar{X}_{c1} + \bar{X}_{c2} + \ldots + \bar{X}_{cn}}{n}$$

where $\bar{\bar{X}}$ is the estimate of the population mean

$\bar{X}_{c1}$, $\bar{X}_{c2}$ etc. are the mean values for each cluster

n is the number of clusters sampled

3. the estimate of the standard deviation of the cluster means if given by:

$$s_{c\bar{X}} = \sqrt{\frac{\sum (\bar{X}_c - \bar{\bar{X}})^2}{n-1}}$$

where $\bar{X}_c$ and $\bar{\bar{X}}$ are the same as in (2) above.

4. the standard error of the estimated population mean, $\bar{X}$, may be obtained by the following relatively simple formula when no apportionment of error due to subsampling is desired. The overall cluster standard error is estimated by:

$$s_{c\bar{x}} = \frac{s_{c\bar{x}}}{\sqrt{n}} \sqrt{1 - \left[\frac{n \cdot m}{N \cdot M}\right]}$$

where n is the number of clusters sampled

m is the number of sub-plots sampled in each cluster

N is the total number of clusters possible in the population

M is the total number of sub-plots possible in a cluster.

The important point in cluster sampling is that the cluster mean is used to calculate the sample statistics. Each cluster may be thought of as a small group of sub-plots, the mean of which is the sampling unit for purposes of statistical estimates. The advantage of cluster sampling is that the variability in the forest is sampled at a greater intensity with clusters than with a single plot. Also the field crews can spend more productive time in sampling and less time on non-productive travel time. Cluster sampling has been applied very successfully in various parts of the world, but specially in tropical forests where access is extremely difficult.

Numerical example of cluster sampling

We will carry out an example of cluster sampling using hypothetical data and applying the formulas given above.

A forest is to be sampled by the cluster method with each cluster consisting of 6 sub-plots (m). Of the total number of possible clusters (N = 100), four clusters (n) are to be measured for total volume. The arrangement and volume values for each sub-plot in each cluster is shown in Figure 7.8.

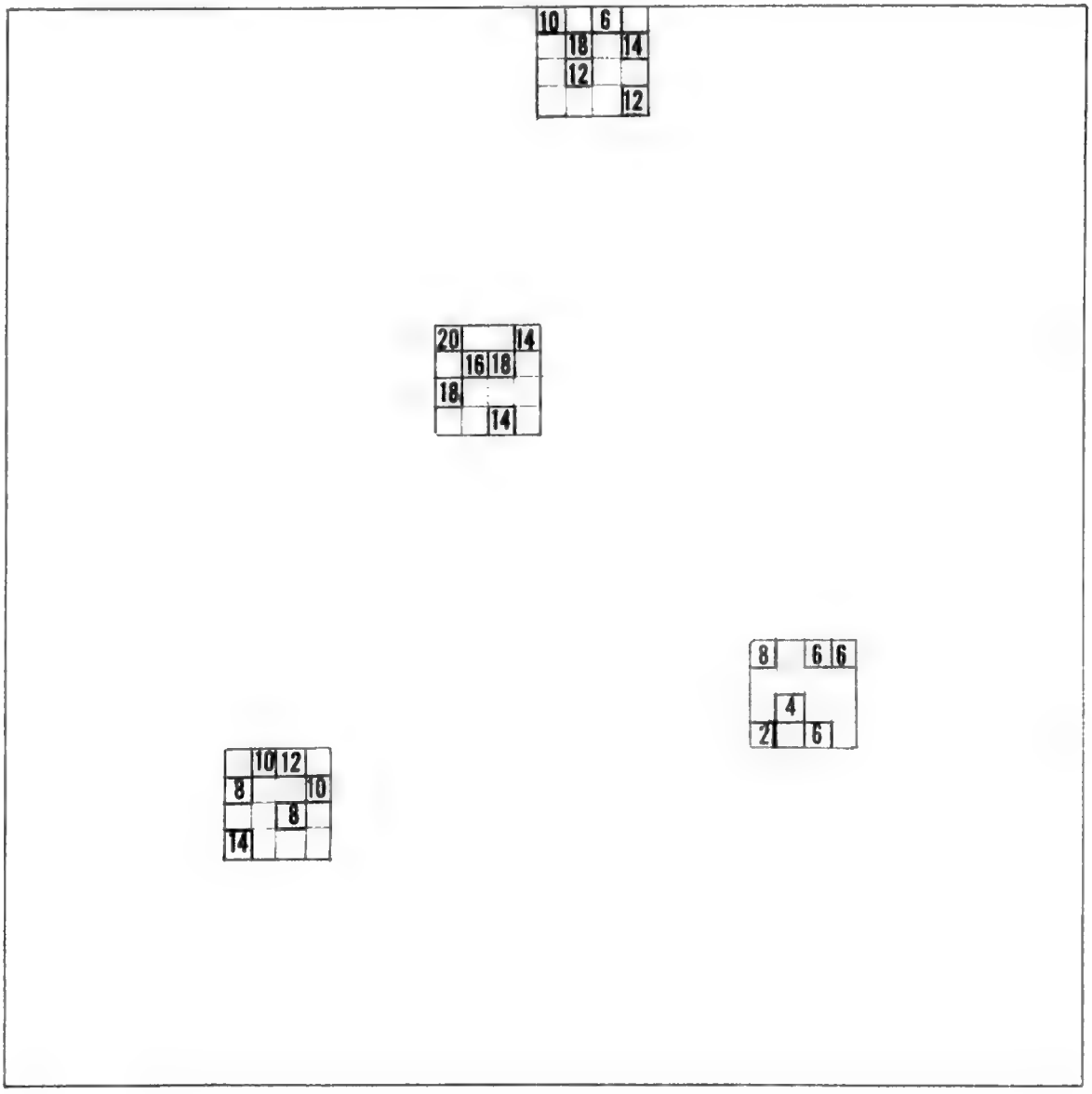


Figure 7.8 Four clusters chosen at random out of a population of 100 clusters. Values shown are in hundreds of board feet per acre for each subplot.

Statistical data for this example are:

Cluster No.	Sub-plot volume	Cluster mean
	——— 00's board feet ———	
1	10, 12, 18 6, 14, 12	12.00
2	20, 18, 16 14, 18, 14	16.70
3	8, 6, 6, 4, 2, 6	5.30
4	10, 12, 8 14, 8, 10	10.30

1. Cluster mean volume derived from sub-plots

$$\bar{\bar{X}} = \frac{(\bar{X}_{c1} + \bar{X}_{c2} + \bar{X}_{c3} + \bar{X}_{c4})}{n}$$

$$= \frac{44.30}{4}$$

$$= 11.08$$

2. Standard deviation of cluster means

$$S_{c\bar{x}} = \sqrt{\frac{\sum (\bar{X}_c - \bar{\bar{X}})^2}{n-1}}$$

$$= \sqrt{\frac{66.4476}{3}}$$

$$= \sqrt{22.1492}$$

$$= \pm 4.706$$

3. Standard error of the mean for the population

$$S_{c\bar{\bar{x}}} = \frac{S_{c\bar{x}}}{\sqrt{n}} \cdot \sqrt{1 - \frac{n \cdot m}{N \cdot M}}$$

$$= \pm \frac{4.706}{\sqrt{4}} \cdot \sqrt{1 - \frac{4 \cdot 6}{100 \cdot 16}}$$

$$= \pm 2.352\ (0.992)$$

$$= \pm 2.333$$

The example of cluster sampling has been made very simple for instructional purposes. In practice, a great number of clusters are established in a forest inventory. A Pre-Investment Survey of Forest Resources project, funded by the United Nations Food and Agriculture Organization (UN/FAO), established 380 clusters, each having eight sub-plots, in the decidous tropical forests of Central India. A number of clusters was measured from each base camp and the savings of travel time were considerable. In the mountainous area of North India, the number of sub-plots in a cluster was reduced to three because of extremely difficult terrain which increased the travel time to and from a cluster.

Cluster sampling is best used for inventories of large areas in which the sampling intensity is low. The number of sub-plots in a cluster is chosen so that the field crew can complete the required measurements on all sub-plots in one day and return to base.

Multiphase Sampling

Multiphase sampling is a devise used in many forest inventory applications. The procedure used in the survey of National Forests in the United States will illustrate the method.

The first phase is to examine and classify a large number of points on aerial photographs. Each point is classified as either (1) forest or (2) non-forest according to set of criteria established for the survey. From this classification list, the forest points are used as the basis for drawing a random or systematic sample of points which will be interpreted on the aerial photographs for forest type, stand height, density and other required characteristics.

Thus, the second phase of sampling is taken from the points making up the first phase. The ratio of interpretation points to forest points may be 1:4 so that one-fourth of the forest points is chosen for more detailed examination.

In the third phase, a proportion of the interpretation points is chosen and will be used as field plots on which the normal inventory data are measured and recorded. The ratio of field plots to interpretation points may be 1:3; therefore one-third of the interpretration points becomes the plot sample. High-cost samples (field plots) are measured on a small proportion of the whole area and adjustments to the photo classifications are made on the basis of the field results.

The sequence in a multiphase sample of this type is illustrated in Figure 7.9.

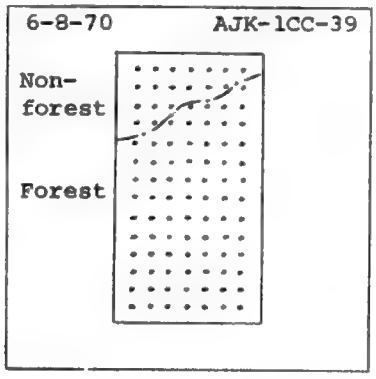


Figure 7.9 A. Dot overlay for initial classification into "Non-forest" and "Forest"

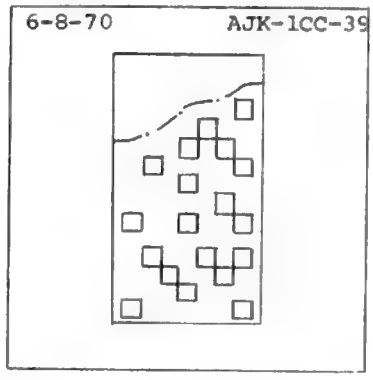


Figure 7.9 B. Random sample of "Forest" points for photo-interpretation.

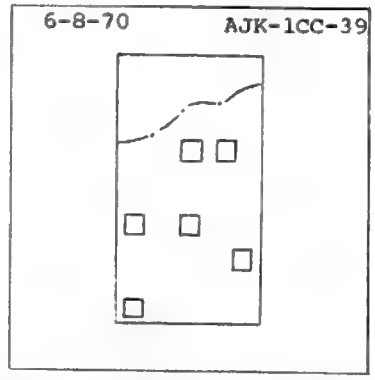


Figure7.9 C. Random sample of interpretation for field plot establishment.

Figure 7.9 Sequence of multiphase sampling

Multistage sampling

A sampling system that is even more sophisticated than the multiphase system is the multistage. Instead of double sampling, as in the case of multiphase sampling, a population is made up of a number of primary sampling units; each primary unit is composed of a number of secondary units. This constitutes a two-stage system. The secondary sampling units may be further broken down into smaller tertiary units to form a three-stage system.

Multi-stage sampling concentrates the measurement work rather than having it spread all over the whole forest. When access to the forested areas is limited by an inadequate road network, the time involved in getting to each of the secondary sampling units is reduced.

The statistical analysis of multistage sampling assumes that one knows analysis of variance techniques. Since this becomes rather complicated, no numerical example will be given but students are referred to Husch, Miller and Beers (1972), Cochran (1963) and to Loetsch and Haller (1964) for complete examples of multistage sampling.

To illustrate multistage sampling, let us assume that we have to obtain inventory data over a large forested area in which roads and other transportation networks are scarce. First, the area to be inventoried is divided into equal-size blocks (primary units); from these blocks, a pre-determined number is chosen at random. Each block is composed of a number of equal-sized secondary sampling units; depending on the design of the survey, each primary block may have two, four, sixteen or some other number of secondary units. A random sample is made of the secondary units within each block which makes up the primary sample. At this point, we have a two-stage sampling design. We could go further and subdivide each secondary sampling unit into smaller units (tertiary sampling units) and obtain a random sample from it to develop a three-stage design.

The random selection of sampling units at all stages is required to obtain unbiased estimates of means and standard errors. Sometimes, the final stage consists of systematically-spaced sampling units; these may be clusters of plots around a central point or plots evenly-spaced along randomly chosen lines. In such a scheme, the cluster of plots or the total number of plots along a line is considered as the sampling unit and the calculation of within-cluster (or within-line) variation is not statistically valid.

Chapter 8

REGRESSION AND CORRELATION

Introduction

The term "regression" was originally used by F. Galton who reported in 1889 that the tendency in human beings was to exhibit the same general peculiarities as their kinsmen but to a lesser degree. Galton's work was expanded by Pearson in 1903 and he showed that tall fathers tend to have tall sons, but the average height of the sons is less than that of the fathers. There is a regression or a going back of the sons' heights to the average height of all men. Nowadays, we use the work regression to depict the relationship between two variables - one independent and the other dependent - whether the relationship is a straight line (linear regression) or is curved (curvilinear regression). For the present, we will concentrate on linear regression; later on, we will examine curvilinear regression in its simplest form.

Mathematical expression of linear regression

Although it is possible to establish a straight line by eye which will show the relation between two variables, this method is not very accurate. Certainly, we could draw a straight line which approximated the trend and we could balance the points about this line just as we balanced a freehand curve in Chapter 1. However, we could not be sure that the line was in the right place. So we resort to a mathematical analysis of the data to establish a formula for the relationship.

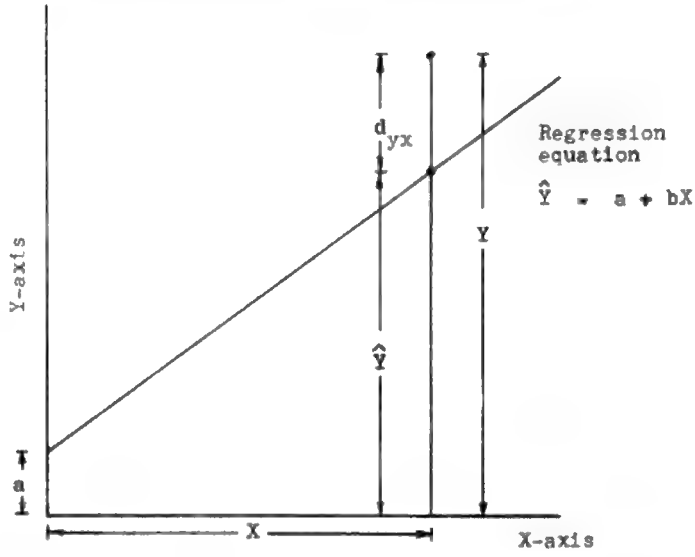


Figure 8.1 - Diagram showing the relation between X, Y, $\hat{Y}$ and d_{yx}.

A straight line has the general formula: -

$$\hat{Y} = a + bX \qquad (8.1)$$

where $\hat{Y}$ = estimated value of the dependent variable

a = value of the Y intercept when X = 0

b = a coefficient establishing the slope of the line

X = a measured value of the independent variable.

The method we shall use is that called the method of least squares. By this we mean that the sum of the squared deviations from the points to the straight line is the least possible. For any other position of the straight line, the sum of squares will be larger.

In order to be consistent, let us learn the symbols which are used in regression analysis. Refer to Figure 8.1 as you read the symbols,

1. X = a measured value of the independent variable
2. Y = a measured value of the dependent variable
3. $\hat{Y}$ = an estimated value of the dependent variable resulting from a measured value of the independent variable.
4. d_{yx} = the deviation of Y on X or the difference between the measured and estimated value of the dependent variable. d_{yx} is sometimes given as 'y'.
5. s_{yx} = standard deviation from regression; also called standard error of estimate. s_{yx} is not shown in Figure 8.1 but will be explained later.

From Figure 8.1, you can see that any measured value of the dependent variable is made up of two components, $\hat{Y}$ and d_{yx} or $Y = \hat{Y} + d_{yx}$.

There are two ways of obtaining the formula for a straight line by the method of least squares and we shall examine both of them.

Method 1, using normal equations

The objective in a least squares solution is to establish values for the coefficients 'a' and 'b' in the general formula. If these are known, and an arbitrary value for X is inserted into the equation, the estimated value of the dependent variable ($\hat{Y}$) can be easily determined. The procedure involves the use of two normal equations and of solving these equations simultaneously. The two normal equations are: -

$$\text{I:} \quad \Sigma Y = a \Sigma f + b \Sigma X \tag{8.2}$$

$$\text{II:} \quad \Sigma XY = a \Sigma X + b \Sigma X^2$$

The procedure for solving two equations simultaneously will be explained in the numerical example below.

Example of least squares method

Basal area per acre is sometimes used to estimate volume in a stand and our problem is to determine an equation for linear regression between the two. A rough plotting of the points indicates that the relationship is linear, so we can start with that assumption. Data were taken in natural oak stands with the independent variable being basal area in square feet per acre for trees 10" dbh and over and the dependent variable being board foot volume by the International 1/4" rule. Fifteen plots were measured and the data have been modified slightly to allow easier calculation. Table 8.1 shows the headings which are necessary to solve our equations.

Table 8.1 - Measurements of X and Y and other information required to determine the equation for linear regression.

	X sq.ft.	Y fbm (000's)	XY	X^2
	18	0.7	12.6	324
	20	1.1	22.0	400
	22	1.3	28.6	484
	25	1.7	42.5	625
	35	2.4	84.0	1225
	42	3.5	147.0	1764
	45	3.9	175.5	2025
	49	4.1	200.9	2401
	53	4.4	233.2	2809
	58	5.1	295.8	3364
	65	5.3	344.5	4225
	72	6.3	453.6	5184
	76	6.2	471.2	5776
	84	6.5	546.0	7056
	86	7.5	645.0	7396
Σ	750	60.0	3702.4	45058
Mean	50	4.0		

The totals must be substituted into the two normal equations given on page 62. They are repeated here so the whole operation may be easily followed.

Normal equations: $\Sigma Y = a \Sigma f + b \Sigma X$

$\Sigma XY = a \Sigma X + b \Sigma X^2$

substituting known values in these equations, we have: -

$$\text{I:} \quad 60 = 15a + 750b$$

$$\text{II:} \quad 3702.4 = 750a + 45058b$$

From this point, the procedure is: -

a. divide Equation I by 15 to bring the value of 'a' to 1.0

b. divide Equation II by 750 to bring the value of 'a' to 1.0

$$\text{I} \div 15: \quad 4.000 = a + 50.000b$$

$$\text{II} \div 750: \quad 4.936 = a + 60.077b$$

c. subtract the two equations in part (b). We have: -

$$-0.936 = -10.077b$$

d. determine the value of 'b'

$$b = \frac{0.936}{10.077}$$

$$= 0.093$$

e. having found the value of 'b', substitute this value into one of the equations which has both 'a' and 'b'. We can use the first equation in part (b) above: -

$$4.000 = a + 50(0.093)$$

$$= a + 4.650$$

[Transpose and change signs] $a = -4.650 + 4.000$

$$= -0.650$$

f. having determined the values for both 'a' and 'b', we may use these in the general equation for a straight line. The equation which fits these data, and no other, is: -

$$\hat{Y} = -0.650 + 0.093X$$

g. for any value of X, a value of Y can be estimated. Set up a table showing various values of X and calculate the $\hat{Y}$ values.

X	$\hat{Y}$
10	0.28
20	1.21
30	2.14
40	3.07
60	4.93
80	6.79

h. plot these points on a graph and draw a line connecting them. You will see that the line passes directly through each point. Any three points will establish a straight line, and it is never safe to use only two, in case an error in either computation or plotting has occurred. The completed graph with the regression line and original points is shown in Figure 8.2.

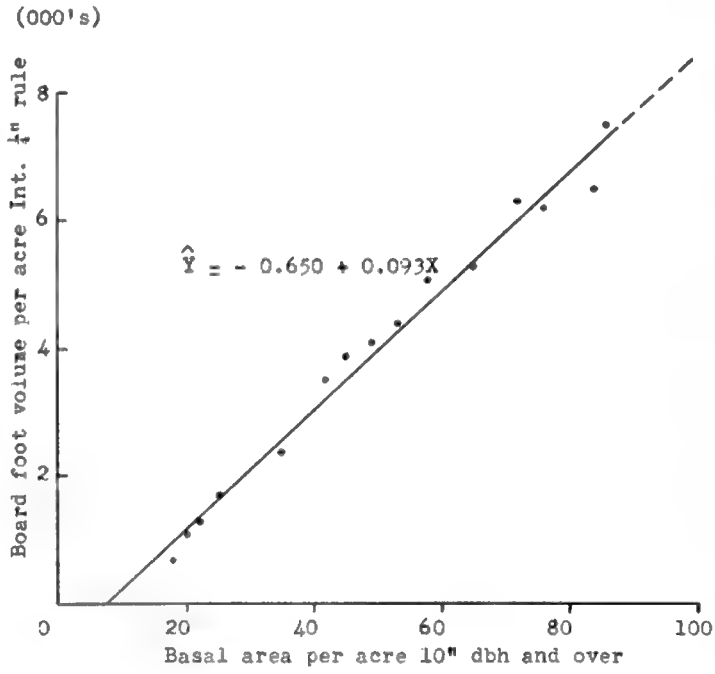


Figure 8.2 - Relation between basal area per acre and volume in board feet for natural oak stands. Regression equation: $\hat{Y} = -0.650 + 0.093X$.

Extrapolation

The procedure of extending a straight line to include classes of X which were not actually measured is called 'extrapolation'. In Figure 8.2 we could extend the straight line so that an X value of 95 could be used to estimate Y. Some caution must be exercised at this point in case the line is extended beyond the region in which the linear relationship is applicable.

Method 2, using deviations from means

The second method for determining the straight line formula uses deviations from the means and cross-products of the deviations. We shall use the same data as before and you can compare the results.

The formula is taken from analytical geometry: -

$$\hat{Y} - \bar{Y} = b(X - \bar{X}) \qquad (8.3)$$

where $\hat{Y}$ = estimated value of the dependent variable

$\bar{Y}$ = arithmetic mean of the dependent variable

b = slope of the line given by the ratio

$$\frac{\Sigma xy}{\Sigma x^2}$$

X = any value of the independent variable

$\bar{X}$ = arithmetic mean of the independent variable.

As we previously learned, lower case letters denote deviations from the mean, so: -

$$x = (X - \bar{X})$$

$$y = (Y - \bar{Y})$$

$$xy = (X - \bar{X})(Y - \bar{Y})$$

$$x^2 = (X - \bar{X})^2$$

Table 8.2 gives the necessary column headings and data for completing the solution of the formula.

Table 8.2 - Original measurements of X and Y, deviations and other data for determining a straight line formula by Method 2.

	X	Y	x	y	x^2	xy
	18	0.7	- 32	- 3.3	1024	+ 105.6
	20	1.1	- 30	- 2.9	900	+ 87.0
	22	1.3	- 28	- 2.7	784	+ 75.6
	25	1.7	- 25	- 2.3	625	+ 57.5
	35	2.4	- 15	- 1.6	225	+ 24.0
	42	3.5	- 8	- 0.5	64	+ 4.0
	45	3.9	- 5	- 0.1	25	+ 0.5
	49	4.1	- 1	+ 0.1	1	- 0.1
	53	4.4	+ 3	+ 0.4	9	+ 1.2
	58	5.1	+ 8	+ 1.1	64	+ 8.8
	65	5.3	+ 15	+ 1.3	225	+ 19.5
	72	6.3	+ 22	+ 2.3	484	+ 50.6
	76	6.2	+ 26	+ 2.2	676	+ 57.2
	84	6.5	+ 34	+ 2.5	1156	+ 85.0
	86	7.5	+ 36	+ 3.5	1296	+ 126.0
Sums:	750	60.0	0	0	7558	+ 702.6
Mean:	50	4.0				

$$b = \frac{702.6}{7558}$$

$$= 0.093$$

From the formula: $\hat{Y} - 4.0 = 0.093\ (X - 50)$

$$= 0.093X - 4.650$$

$$\hat{Y} = - 4.650 + 4.000 + 0.093X$$

$$= - 0.650 + 0.093X$$

You will see that the final formula is the same by both methods, so it really does not matter which one you use. There is, perhaps, less chance for making computational errors in Method 2 because the figures do not become so complicated. Whichever method is used, the computations must be done carefully and neatly.

If you use a desk calculator to obtain the sum of squares and other data, you might try the following method for obtaining Σx^2: -

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \qquad (8.4)$$

The ΣX^2 can be cumulated on the calculator quite easily; the term $\frac{(\Sigma X)^2}{N}$ is a correction factor to reduce the ΣX^2. Using the same data, the ΣX^2 is. -

$$\Sigma x^2 = 45058 - \frac{(750)^2}{15}$$

$$= 45058 - \frac{562500}{15}$$

$$= 45058 - 37500$$

$$= 7558$$

Standard deviation from regression

In the calculation of standard deviation (Chapter 5) we took deviations from a fixed mean and calculated the variability of the items in the sample around the mean. We used the formula $s = \sqrt{\frac{\Sigma x^2}{N - 1}}$

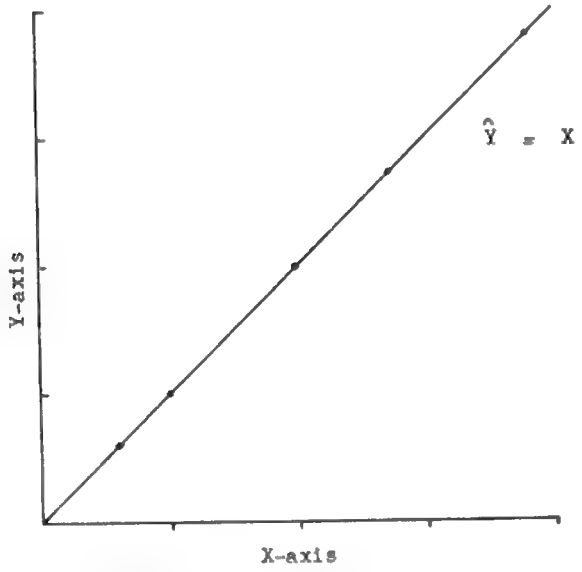


Figure 8.3 - Graph of $\hat{Y} = X$.

A similar treatment can be given in the case of deviations from regression except that we are now dealing with a moving mean - the regression line - rather than with a fixed mean. The size of the deviations from the regression line is an indication of the variability of the sampling units. If we had a relationship such that $X = Y$, the resulting regression equation would be $\hat{Y} = X$; each plotted point would fall exactly on the regression line and each deviation $(Y - \hat{Y})$ would be 0. The total variation of the points about the line would obviously be 0. Figure 8.3 illustrates this special case.

The most widely-used term for describing the variability of items about a moving mean is standard deviation from regression; it permits us to think of the deviations in the same light as those for standard deviation. An expression found in some texts to describe the variability of items about a moving mean is standard error of estimate; this could be confused with standard error of the mean, and the preferred term is standard deviation from regression.

The formula for standard deviation from regression is: -

$$s_{yx} = \sqrt{\frac{\Sigma d_{yx}^2}{N - 2}} \qquad (8.5)$$

where s_{yx} = standard deviation from regression

d_{yx}^2 = the sum of the squared deviations, taken in the Y-direction, of the points from the regression line. $d_{yx} = (Y - \bar{Y})$

$N - 2$ = degrees of freedom.

Note the difference in the degrees of freedom for standard deviation from regression (N - 2) from those for standard deviation (N - 1). The reduction in degrees of freedom from (N - 1) to (N - 2) is attributed to the fact that the slope of the regression line (b) is a constant. N - 2 is valid for degrees of freedom for linear regression, not for curvilinear regression where d.f. = N - 3 or N - 4 etc. depending upon the complexity of the curve. One additional note is that if the points making up the curve represent the total number of items in the population, as opposed to the points being a sample from a population, the divisor in the standard deviation from regression formula would be N instead of N - 2. In most forestry applications, a sample is taken to represent the population, rather than all items being measured; therefore, the usual divisor will be N - 2.

To illustrate the calculation of standard deviation from regression, we will cite the example in which basal area in square feet per acre for trees 10" dbh and over is used to

estimate board foot volume in natural oak stands. Table 8.3 shows the necessary column headings for the calculation of s_{yx}. $\hat{Y}$ in the table below is obtained by substituting particular values of X in the regression equation $\hat{Y} = - 0.650 + 0.093X$.

Table 8.3 - Calculation of s_{yx} in which X = basal area in square feet per acre for trees 10" dbh and over, Y = board feet per acre International 1/4" rule (1000).

	X	Y	$\hat{Y}$	d_{yx}	d_{yx}^2
	18	0.7	1.02	- 0.32	0.1024
	20	1.1	1.21	- 0.11	0.0121
	22	1.3	1.39	- 0.09	0.0081
	25	1.7	1.67	+ 0.03	0.0009
	35	2.4	2.60	- 0.20	0.0400
	42	3.5	3.25	+ 0.25	0.0625
	45	3.9	3.53	+ 0.37	0.1369
	49	4.1	3.90	+ 0.20	0.0400
	53	4.4	4.27	+ 0.13	0.0169
	58	5.1	4.73	+ 0.37	0.1369
	65	5.3	5.37	- 0.07	0.0049
	72	6.3	6.04	+ 0.26	0.0676
	76	6.2	6.41	- 0.21	0.0441
	84	6.5	7.15	- 0.65	0.4225
	86	7.5	7.34	+ 0.16	0.0256
Sums:	750	60.0			1.1214

$$s_{yx} = \sqrt{\frac{1.1214}{15 - 2}}$$

$$= \sqrt{\frac{1.1214}{13}}$$

$$= \sqrt{0.0862}$$

$$= \pm 0.293$$

The limits of the regression line $\pm s_{yx}$ are shown in Figure 8.4.

The interpretation of the statistic s_{yx} is that it is a measure of the variability of the measured Y values which is not accounted for by the linear relationship with X. As we pointed out previously, if the relationship between X and Y were perfect, there would be no variability in $(Y - \hat{Y})$ and hence $s_{yx} = 0$. s_{yx} is the inability of the measured Y values to agree with the equation $\hat{Y} = - 0.650 + 0.093X$ that is being recognized in the standard deviation from regression.

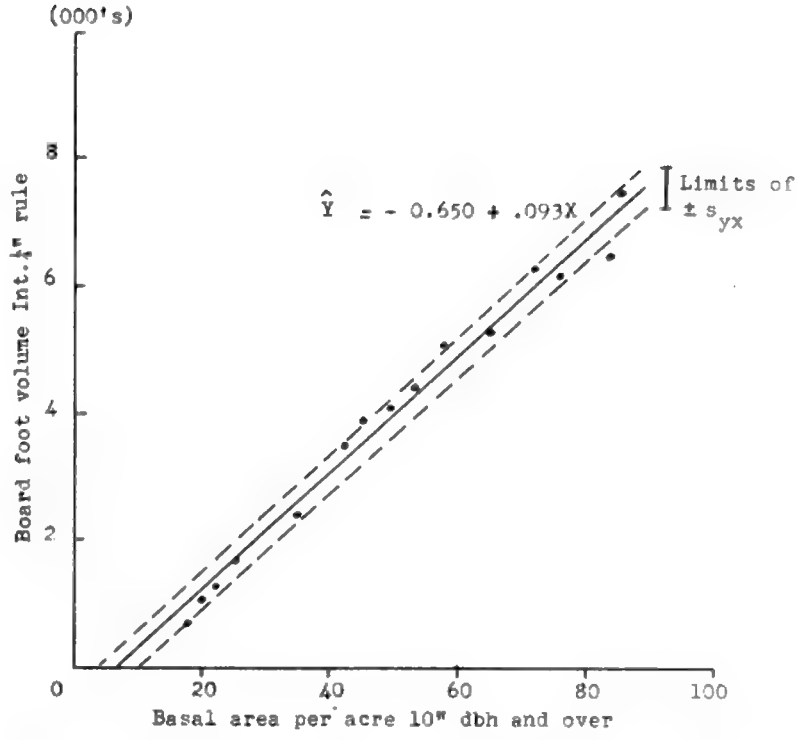


Figure 8.4 - Regression equation $\hat{Y} = - 0.650 + 0.093X$ with limits of $\pm s_{yx}$.

Standard error of the regression coefficient

The regression coefficient 'b' is an estimate since the original values of X and Y are sampling units drawn from an infinite population. Therefore, our calculated regression coefficient 'b' is an

estimate of a population regression coefficient 'β' and is subject to sampling variation. The statistic, standard error of the regression coefficient is: -

$$s_b = \frac{s_{yx}}{\sqrt{\Sigma x^2}} \qquad (8.6)$$

where s_b = standard error of the regression coefficient

s_{yx} = standard deviation from regression

Σx^2 = the sum of the squared deviations or $\Sigma(X - \bar{X})^2$.

In the example,

$$s_b = \frac{0.293}{\sqrt{7558}}$$

$$= \frac{0.293}{86.9}$$

$$= \pm 0.00337$$

The population regression coefficient 'β' will lie within the range $b \pm t_{.05}s_b$ at the 5% level or -

$$b - t_{.05}s_b < \beta < b + t_{.05}s_b$$

The value of 't' for d.f. = 13 at the 5% level is 2.160, so we have: -

$$0.093 - 2.160(0.00337) < \beta < 0.093 + 2.160(0.00337)$$

$$0.093 - 0.00728 < \beta < 0.093 + 0.00728$$

$$0.086 < \beta < 0.100$$

The population parameter 'β' will lie within the limits of 0.086 to 0.100 unless one chance in 20 has occurred.

Testing the significance of the regression coefficient

A test of whether the regression coefficient is actually significant or not can be obtained from the formula: -

$$t = \frac{b}{s_b} \qquad (8.7)$$

In our case, we have b = 0.093 and s_b = 0.00337, so -

$$t = \frac{0.093}{0.00337}$$

$$= 27.6\ **$$

The tabulated value of 't' for d.f. = 13 at the 5% level is 2.160 and at the 1% level is 3.012, so our calculated 't' is highly significant and is given a double asterisk. It means that a larger value of 't' would occur by chance less than one time in 100. Actually, the chances are much less than that, but we do not have tabulated values of 't' for less than the 1% level on page 40. Some texts publish values to the 0.01% level. What we are, in fact, measuring is whether the regression equation characterized by 'b' is significantly different from using the average of the Y values ($\bar{Y}$) to portray the relationship between Y and X. You can see from Figure 8.4 that there is a very strong relationship between basal area in square feet per acre for trees 10" dbh and over and board foot volume per acre in natural oak stands. It is no surprise that this should be so because the two variables are quite intimately related. It would be surprising if the regression coefficient were not highly significant.

Computation of the error in predicting $\hat{Y}$, standard error of $\hat{Y}$

When we use a value of X to predict $\hat{Y}$ through the regression equation, our predicted $\hat{Y}$ is subject to some error because the original values of X and Y were samples and subject to their own error. It is of considerable interest to know, not only the value of a predicted $\hat{Y}$ but also how good our estimate is. What are the chances that our predicted values are within acceptable limits? How certain can we be that the predicted values are satisfactory? We have two alternatives in calculating error of a predicted $\hat{Y}$, symbolized by $s_{\hat{Y}}$. The first is when we assume that $X \neq \bar{X}$ and that $\hat{Y}$ is subject to variation. The formula for this condition is: -

$$s_{\hat{Y}} = s_{yx}\sqrt{\frac{1}{N} + \frac{(X - \bar{X})^2}{\Sigma x^2}} \qquad (8.8)$$

The second condition is when predictions are being made for individual values of the independent variable. In this case, Formula 8.8 is modified to: -

$$s_{\hat{Y}} = s_{yx}\sqrt{1 + \frac{1}{N} + \frac{(X - \bar{X})^2}{\Sigma x^2}} \qquad (8.9)$$

The symbols used in Formulas 8.8 and 8.9 have been explained previously. The deviation $(X - \bar{X})$ is the deviation of the chosen X value from the mean of the original X values, not from the mean of the values used to compute $s_{\hat{Y}}$.

If we choose the 5% level of probability as our reference, we will have the following data to put into the formula: -

$$\hat{Y} = -0.650 + 0.093X$$

$$N = 15$$

$$\Sigma x^2 = 7558$$

$$s_{yx} = \pm 0.293$$

$$t_{.05} = 2.160$$

Let us use X = 10, 20, 30, 40, 50, 60, 70 and 80 square feet as the points for which to compute $s_{\hat{Y}}$. For each of these X values, we can compute a $\hat{Y}$ from the regression formula using Formula 8.8. Table 8.4 shows the headings and data.

Table 8.4 - Data for computing the lower and upper fiducial limits of $s_{\hat{Y}}$.

X	Y	$(X-\bar{X})$	$(X-\bar{X})^2$	$\frac{(X-\bar{X})^2}{\Sigma x^2}$	$\sqrt{\frac{1}{N} + \frac{(X-\bar{X})^2}{\Sigma x^2}}$	$s_{\hat{Y}}$	$t_{s_{\hat{Y}}}$	Limits	
								Lower	Upper
10	0.28	- 40	1600	0.212	0.528	0.155	0.335	-	0.615
20	1.21	- 30	900	0.119	0.430	0.126	0.272	0.938	1.482
30	2.14	- 20	400	0.053	0.346	0.101	0.218	1.922	2.358
40	3.07	- 10	100	0.013	0.283	0.083	0.179	2.891	3.249
50	4.00	0	0	0	0	0	0	4.000	4.000
60	4.93	10	100	0.013	0.283	0.083	0.179	4.751	5.109
70	5.86	20	400	0.053	0.346	0.101	0.218	5.642	6.078
80	6.79	30	900	0.119	0.430	0.126	0.272	6.518	7.062

Figure 8.5 shows the original points, the regression line and the lower and upper limits of the 5% level of $s_{\hat{Y}}$.

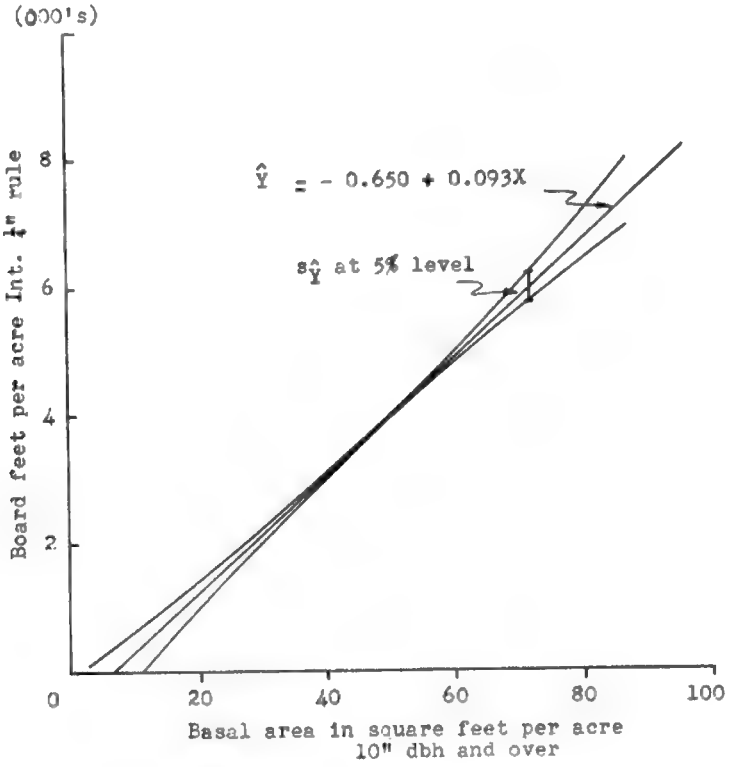


Figure 8.5 - Graph of $\hat{Y} = -0.650 + 0.093X$ with upper and lower limits of $s_{\hat{Y}}$ at the 5% level.

Curvilinear regression

The determination of a regression equation to fit data which exhibit a non-linear relationship becomes considerably more complicated. In this beginning treatment of statistics, we shall be concerned only with the solution of a second degree curve which has the general formula: -

$$\hat{Y} = a + bX + cX^2 \qquad (8.10)$$

Notice that we still have only two variables, X and Y, but that the independent variable is shown as its second power (X^2) in addition to its normal form. Formula 8.10 is the general equation for a parabolic relation.

In linear regression, we used two normal equations for determining the coefficients 'a' and 'b'; in curvilinear regression, we use three normal equations which are: -

$$\begin{aligned} \Sigma Y &= a \Sigma f + b \Sigma X + c \Sigma X^2 \\ \Sigma XY &= a \Sigma X + b \Sigma X^2 + c \Sigma X^3 \\ \Sigma X^2 Y &= a \Sigma X^2 + b \Sigma X^3 + c \Sigma X^4 \end{aligned} \qquad (8.11)$$

The relationship between diameter in inches and volume in board feet follows a second degree curve; in fact, the volume in board feet by the International 1/8" rule for 16-foot has the formula: -

$$\hat{V} = -1.40 - 1.52D + 0.88D^2$$

which corresponds to the general parabolic formula.

The procedure for solving three equations simultaneously will be explained in the following example. No units of measure have been given to X and Y, and the original measurements are assumed to be randomly selected samples from a population of X's. The Y's are measurements related to X. Table 8.6 shows the original measurements.

Table 8.6 - Original measurements of X and Y from a population of X's.

X	Y
2	2
7	2
10	5
14	4
19	5
21	10
22	7
26	11
28	16
32	21
34	22
35	28
37	28
38	34
39	40
40	37

In order to solve the three equations simultaneously, we need the following: -

$$\Sigma X,\ \Sigma Y,\ \Sigma XY,\ \Sigma X^2,\ \Sigma X^2Y,\ \Sigma X^3 \text{ and } \Sigma X^4$$

The sum of squares and sum of cross-products can be obtained quite efficiently on a desk calculator by cumulating the products of the individual squares or cross-products. Table 8.7 shows the complete tabulations to illustrate the method.

Table 8.7 - Data for obtaining sums of squares and sums of cross-products for computing a second degree equation.

	X	Y	XY	X^2	X^2Y	X^3	X^4
	2	2	4	4	8	8	16
	7	2	14	49	98	343	2401
	10	5	50	100	500	1000	10000
	14	4	56	196	784	2744	38416
	19	5	95	361	1805	6859	130321
	21	10	210	441	4410	9261	194481
	22	7	154	484	3388	10648	234256
	26	11	286	676	7436	17576	456976
	28	16	448	784	12544	21952	614656
	32	21	672	1024	21504	32768	1048576
	34	22	748	1156	25432	39304	1336336
	35	28	980	1225	34300	42875	1500625
	37	28	1036	1369	38332	50653	1874161
	38	34	1292	1444	49096	54872	2085136
	39	40	1560	1521	60840	59319	2313441
	40	37	1480	1600	59200	64000	2560000
Sums:	404	272	9085	12434	319677	414182	14399798

Having obtained the sum of each column in Table 8.7, the appropriate figures are substituted into the normal equations.

$$\text{I:} \quad 272 = 16a + 404b + 12434c$$

$$\text{II:} \quad 9085 = 404a + 12434b + 414182c$$

$$\text{III:} \quad 319677 = 12434a + 414182b + 14399798c$$

a. divide each equation by its coefficient for 'a'

1. I ÷ 16: $17.00 = a + 25.25b + 777.12c$

2. II ÷ 404: $22.49 = a + 30.78b + 1025.20c$

3. III ÷ 12434: $25.71 = a + 33.31b + 1158.10c$

b. subtract Equation a(1) from a(2)

$$\begin{array}{rrrrrrr} 22.49 & = & a & + & 30.78b & + & 1025.20c \\ 17.00 & = & a & + & 25.25b & + & 777.12c \\ \hline \end{array}$$

1. $5.49 = 5.53b + 248.08c$

c. subtract Equation a(2) from a(3)

$$\begin{array}{rrrrrrr} 25.71 & = & a & + & 33.31b & + & 1158.10c \\ 22.49 & = & a & + & 30.78b & + & 1025.20c \\ \hline \end{array}$$

1. $3.22 = 2.53b + 132.90c$

d. divide b(1) and c(1) by their respective coefficients for 'b'

1. b(1) ÷ 5.53: $0.992 = b + 44.861c$

2. c(1) ÷ 2.53: $1.273 = b + 52.530c$

e. subtract d(1) from d(2)

$$\begin{array}{rrrrrrr} 1.273 & = & b & + & 52.530c \\ 0.992 & = & b & + & 44.861c \\ \hline \end{array}$$

$$0.281 = 7.669c$$

f. determine 'c'

$$7.669c = 0.281$$

$$c = \frac{0.281}{7.669}$$

$$= 0.03664$$

g. substitute c = 0.03664 into an equation containing 'b' and 'c'

c = 0.03664 in Equation d(1)

$$0.9920 = b + 44.861(0.03664)$$

$$= b + 1.6437$$

Transpose and change signs $b = -0.6517$

h. substitute b = - 0.6517 and c = 0.03664 into an equation containing 'a', 'b' and 'c'

b = - 0.6517

c = 0.03664 in Equation a(2)

$$22.49 = a + 30.78(-0.6517) + 1025.20(0.03664)$$

$$= a - 20.0593 + 37.5633$$

Transpose and change signs

$$a = 22.49 + 20.0593 - 37.5633$$

$$= 4.9860$$

The three coefficients are therefore: -

$$a = 4.9860$$

$$b = - 0.6517$$

$$c = 0.03664$$

and the second degree equation becomes: -

$$\hat{Y} = 4.9860 - 0.6517X + 0.03664X^2.$$

The position of the curve can be obtained by substituting various values of X into the final equation and connecting the points. We will need more points than are the minimum requirement for a straight line because the shape of the second degree curve is constantly changing. Table 8.8 gives the estimated values of the dependent variables for X = 0, 10, 20, 30 and 40.

Table 8.8 - Values of $\hat{Y}$ for X = 0, 10, 20, 30 and 40 from $\hat{Y} = 4.9860 - 0.6517X + 0.03664X^2$

X	$\hat{Y}$
0	4.986
10	2.133
20	6.608
30	18.411
40	37.528

These values for $\hat{Y}$, as well as the original points, are plotted and the second degree curve drawn in Figure 8.6.

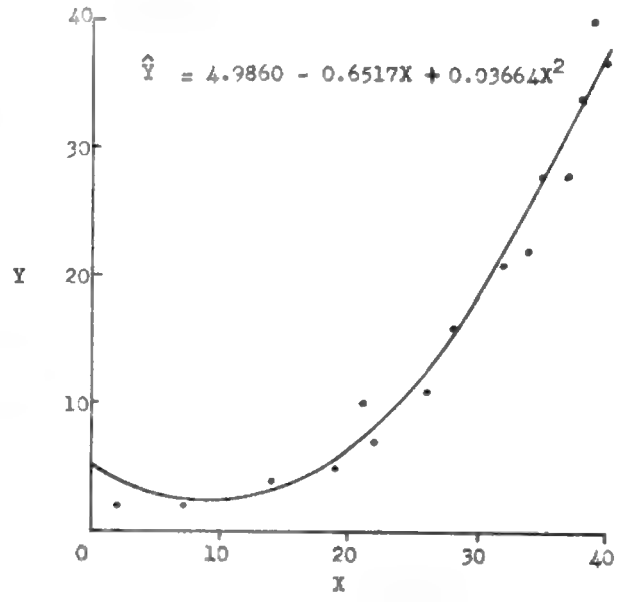


Figure 8.6 - Graph of second degree equation $\hat{Y} = 4.9860 - 0.6517X + 0.03664X^2$.

In a second degree regression curve, the slope of the line changes continually; there is no point in computing the standard error of the regression coefficient as we did for linear regression.

Standard deviation from regression would be interesting in that it would determine, to some extent, whether the data were from a reasonably normal population. The limits of the regression line $\pm s_{yx}$ indicates that approximately 68% of the items would lie within the band if the population were normally distributed. For the curvilinear data, s_{yx} is calculated from Table 8.9.

Table 8.9 - Data for computing standard deviation from the curvilinear regression equation $\hat{Y} = 4.9860 - 0.6517X + 0.03664X^2$.

	X	Y	$\hat{Y}$	d_{yx}	d_{yx}^2
	2	2	3.83	- 1.83	3.3489
	7	2	2.20	- 0.20	0.0400
	10	5	2.13	+ 2.87	8.2369
	14	4	3.03	+ 0.97	0.9409
	19	5	5.83	- 0.83	0.6889
	21	10	7.46	+ 2.54	6.4516
	22	7	8.38	- 1.38	1.9044
	26	11	12.81	- 1.81	3.2761
	28	16	15.46	+ 0.54	0.2916
	32	21	21.65	- 0.65	0.4225
	34	22	25.18	- 3.18	10.1124
	35	28	27.06	+ 0.94	0.8836
	37	28	31.03	- 3.03	9.1809
	38	34	33.13	+ 0.87	0.7569
	39	40	35.30	+ 4.70	22.0900
	40	37	37.54	- 0.54	0.2916
Sums:	404	272			68.9174
Means:	25.25	17			

$$s_{yx} = \sqrt{\frac{\Sigma d_{yx}^2}{N - 3}} \qquad (8.12)$$

$$= \sqrt{\frac{68.9174}{13}}$$

$$= \sqrt{5.301}$$

$$= \pm 2.30$$

Note that d.f. = 13 or N - 3; the explanation of changing d.f. with linear and curvilinear regressions was given on page 66.

Taking $s_{yx} = \pm 2.30$, the upper and lower limits can be plotted and a rough check on normality taken by counting the points which lie within the band. For our sample of N = 16, this was done and 11 out of the 16 were within the limits of the regression curve $\pm s_{yx}$. This amounted to 68.75% - a reasonably good indication of normality considering the small sample of 16 items.

The main interest in curvilinear regression is to determine whether the second degree curve caused a significant reduction in the sum of squares for linear regression. If the curvilinear regression did not cause a significant reduction in the sum of squares, there would be no advantage to calculating it and we might just as well have used linear regression to portray the relationship between Y and X.

In order to arrive at a decision, we must calculate the linear regression and then compare the sums of squares. Without going through the complete procedure in detail, but having followed the example at the beginning of this chapter, the formula for linear regression for the same data is: -

$$\hat{Y} = - 8.098 + 0.994X$$

Σd_{yx}^2 from linear regression must be calculated and these data are given in Table 8.10.

Table 8.10 - Data for calculating s_{yx} from the linear regression equation $\hat{Y} = -8.098 + 0.994X$.

	X	Y	$\hat{Y}$	d_{yx}	d_{yx}^2
	2	2	- 6.11	+ 8.11	65.7721
	7	2	- 1.24	+ 3.24	10.4976
	10	5	+ 1.84	+ 3.16	9.9856
	14	4	+ 5.82	- 1.82	3.3124
	19	5	+ 10.79	- 5.79	33.5241
	21	10	+ 12.78	- 2.78	7.7284
	22	7	+ 13.77	- 6.77	45.8329
	26	11	+ 17.75	- 6.75	45.5625
	28	16	+ 19.73	- 3.73	13.9129
	32	21	+ 23.71	- 2.71	7.3441
	34	22	+ 25.70	- 3.70	13.6900
	35	28	+ 26.70	+ 1.30	1.6900
	37	28	+ 26.68	+ 1.32	1.7424
	38	34	+ 29.67	+ 4.33	18.7489
	39	40	+ 30.67	+ 9.33	87.0489
	40	37	+ 31.66	+ 5.34	28.5156
Sums:	404	272			394.9084
Means:	25.25	17			

$$s_{yx} = \sqrt{\frac{\Sigma d_{yx}^2}{N - 2}}$$

$$= \sqrt{\frac{394.9084}{14}}$$

$$= \sqrt{28.2077}$$

$$= \pm 5.31$$

We have two sums of squares to compare; the first is that from linear regression, 394.9084 with d.f. = 14 and the second is that from curvilinear regression, 68.9174 with d.f. = 13.

At this point, we can introduce a method of comparing different sums of squares. It is the F ratio, discovered by R. A. Fisher, which is: -

$$F = \frac{\text{Mean square of sample means}}{\text{Mean square of individuals}} \qquad (8.13)$$

The mean square is nothing more than the sum of squares divided by the appropriate degrees of freedom. The F ratio has its own distribution and tables are available showing the 5% and 1% levels of probability for different degrees of freedom for the numerator and denominator. See Table A.4 in the Appendix.

In order to make the comparison between the mean square due to linear regression and that due to curvilinear regression, we can set up a convenient table. Table 8.11 gives the appropriate values and the interpretation of the data will follow.

Table 8.11 Test of Significance of Departure from Linear Regression

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Deviation from linear regression	14	394.9084	28.208
Deviation from curvilinear regression	13	68.9174	5.301
Curvilinearity of regression	1	325.991	325.991

$$F = \frac{\text{mean square of deviations from curvilinear regression}}{\text{mean square of curvilinearity of regression}}$$

$$= \frac{325.991}{5.301}$$

$$= 61.496 \quad **$$

The F-table value from Table A 4 (page 151), for d.f. = 1 for the greater mean square and d.f. = 13 for the lesser mean square is 4.67 for P = .05 and 9.07 for P = .01.

The original hypothesis is that the data may be fitted with a linear regression; however, the reduction in the sum of squares caused by the curvilinear regression proves to be significant at the P = .01. Therefore, we reject the hypothesis of linear regression and state that there is a significant curvilinearity in the data as shown by the second degree equation:

$$Y = 4.9860 - 0.6517\ X + 0.03664\ X^2.$$

Transformation of Variables

If the relationship between two variables can be expressed in curvilinear form, there are two main possibilities for the form of the curve. It may be either:

1. parabolic, with the general formula $Y = aX^b$
or 2. exponential with the general formula $Y = ab^X$

If either of these two general equations fits the data, it is possible to extrapolate data by transforming the curved realtionship into straight lines by plotting the equations on either double-log or single-log graph paper.

Two occasions arise in which transformation is often carried out. The first is in connection with a parabolic type of curve which has the general formula: -

$$\hat{Y} = aX^b \qquad (8.14)$$

If this were transformed into logarithmic form, we would have: -

$$\log \hat{Y} = \log a + b \log X$$

which becomes a straight line when both the X and Y axes are scaled in logarithmic units. It would be a straight line because it corresponds in form to the regular linear equation of $\hat{Y} = a + bX$ except that the terms are in logarithmic form.

When using logarithmic forms (base 10) remember that logarithms range from - to + and that:

$10^0 = 1$; therefore log 1 = 0
$10^1 = 10$; therefore log 10 = 1
$10^2 = 100$; therefore log 100 = 2

For numbers less than 1, the log is negative and:

$10^{-1} = 0.1$; therefore log 0.1 = -1
$10^{-2} = 0.01$; therefore log 0.01 = -2
$10^{-3} = 0.001$; therefore log 0.001 = -3

and that the log of number less than 1 is always written so that the fractional part is positive. Thus:

0.610 may be written 6.10×10^{-1}
log 0.610 = log 6.10 + log 10^{-1}
= 0.78533 - 1

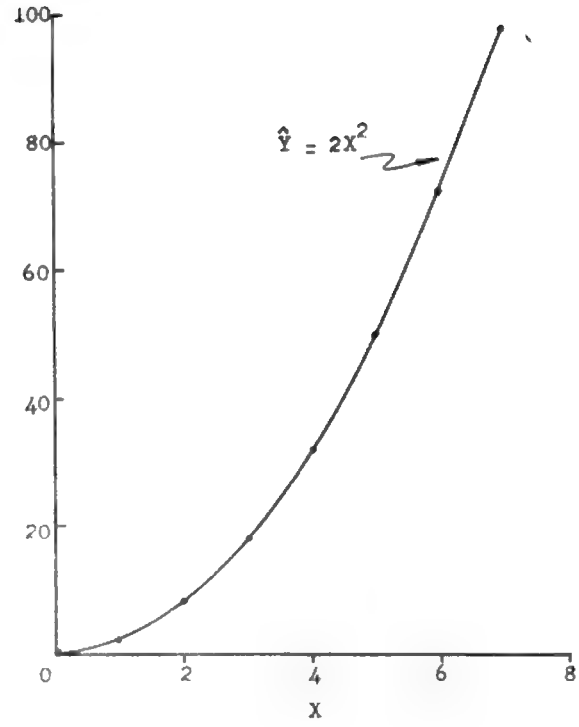


Figure 8.8 - Graph of $\hat{Y} = 2 X^2$.

Take the formula $\hat{Y} = 2 X^2$ for instance; we would determine values for Y for various values of X, plot and draw the resulting graph. For X = 1, 2, 3, 4, 5, 6 and 7, the values of Y = 2, 8, 18, 32, 50, 72 and 98 respectively. These points would plot as shown in Figure 8.8.

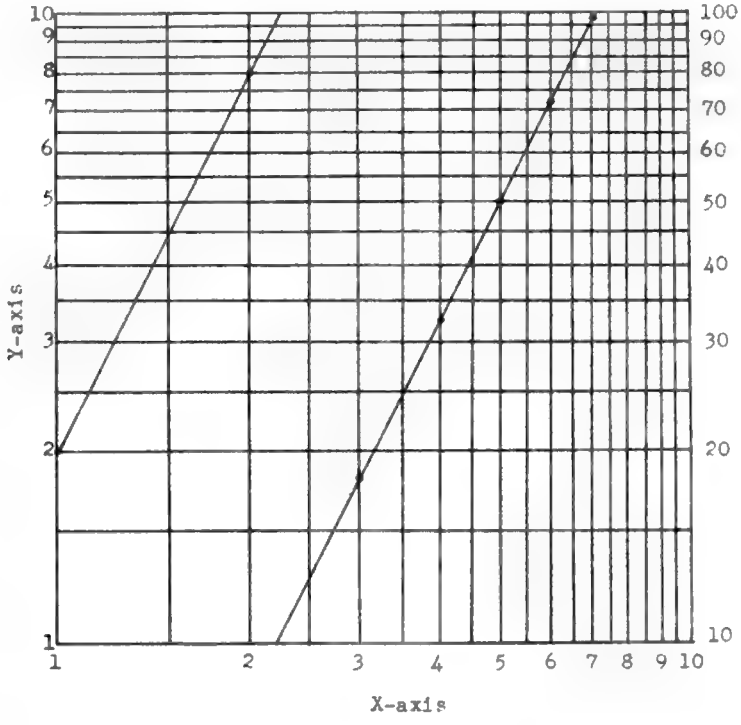


Figure 8.9 - Graph of $\hat{Y} = 2 X^2$ plotted on double log paper.

To transform the variables and plot the points on a logarithmic basis, we have two choices. The first is to look up the appropriate values of X and Y in a table of natural logarithms and to plot the logarithmic values on 10x10 graph paper. This method is not only tedious but impractical. The second choice is to plot the actual values of X and Y on graph paper which has both axes scaled in logarithmic units. The points representing X and Y in the equation $\hat{Y} = 2 X^2$ will fall in a straight line as previously described. Figure 8.9 shows the points and linear relationship between log X and log Y for the parabolic type of equation.

An interesting and instructive advantage to the use of double log paper is the extension of the line to cover high values of X. The equation $\hat{Y} = 2 X^2$ holds true for all values of X, so the same will be true when X and Y are in logarithmic form. If there is only one

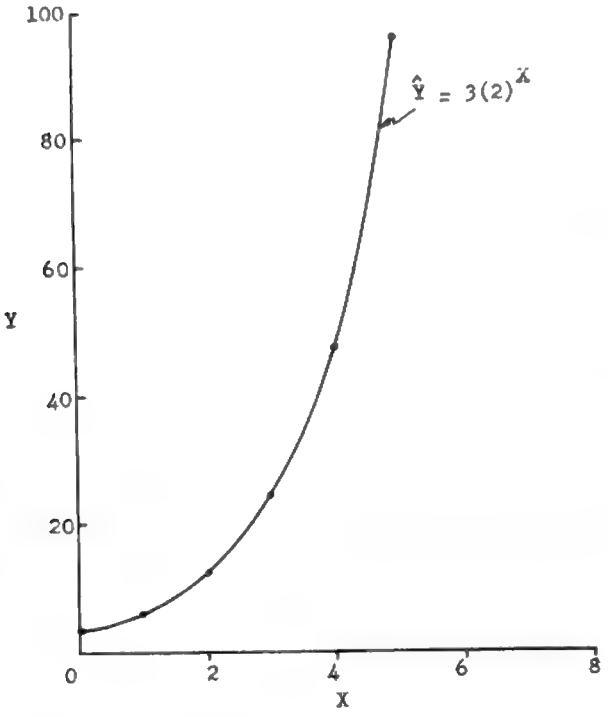


Figure 8.10 - Graph of $\hat{Y} = 3\ (2)^X$.

cycle of Y, as in Figure 8.9, it is possible to shift to two or three cycles by regraduating the Y-axis accordingly and by drawing lines parallel to the original. For the sake of simplicity, only one additional cycle has been added to Figure 8.9 and the new graduations for Y are shown on the right vertical axis.

The second occasion when logarithmic transformation is carried out is when we are dealing with an equation which has the general formula: -

$$\hat{Y} = a\ b^X \qquad (8.15)$$

which is an exponential formula. When transformed to logarithms, this becomes: -

$$\log Y = \log a + X \log b$$

The X in this equation is not in logarithmic form, dictating that the X-axis be graduated into equally-spaced units.

If a = 3 and b = 2, the general exponential formula becomes: -

$$\hat{Y} = 3\ (2)^X$$

and by substituting different values of X, values of Y can be determined. If we set X = 0, 1, 2, 3, 4, and 5, Y becomes 3, 6, 12, 24, 48 and 96 respectively. Plotting these points on 10x10 graph paper results in Figure 8.10.

Now plot the same values on single log graph paper in which only the Y-axis is graduated in logarithmic units. The result will be as in Figure 8.11.

The same extension of the straight line to include higher values of X is possible with the exponential type of transformation as it was with the parabolic.

Correlation

When one measures an attribute of a tree such as diameter at breast height and another, double bark thickness at dbh, the two measurements form a fairly tight band of points when plotted on 10x10 graph paper. One measurement is said to be correlated with the other. The original meaning of the term correlation was that two measurements were related to the same inheritance characteristics; thus, of boys and girls in a family, the heights of the boys was correlated with the heights of the girls because they had the same parents. One was not a cause of the other; there was no cause and effect relationship.

The original idea behind correlation has been expanded to include any two measurements which exhibit a definite relationship

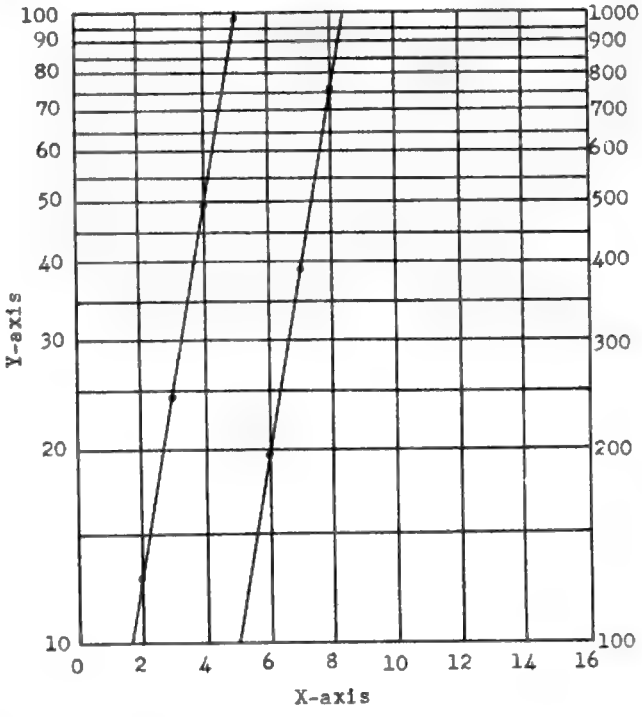


Figure 8.11 - Graph of $\hat{Y} = 3\ (2)^X$ plotted on single log paper.

between them. In this introduction to correlation, we will deal only with linear correlation, and as you might expect, linear regression and linear correlation are quite intimately related.

Before delving into the mathematics of correlation, it would be wise to look at correlation from a general viewpoint. The indicator of the degree of correlation is the coefficient of correlation 'r', which is designed to vary between - 1.0 and ± 1.0. Examples of various coefficients of correlation and the spread of the points are included in Figure 8.12.

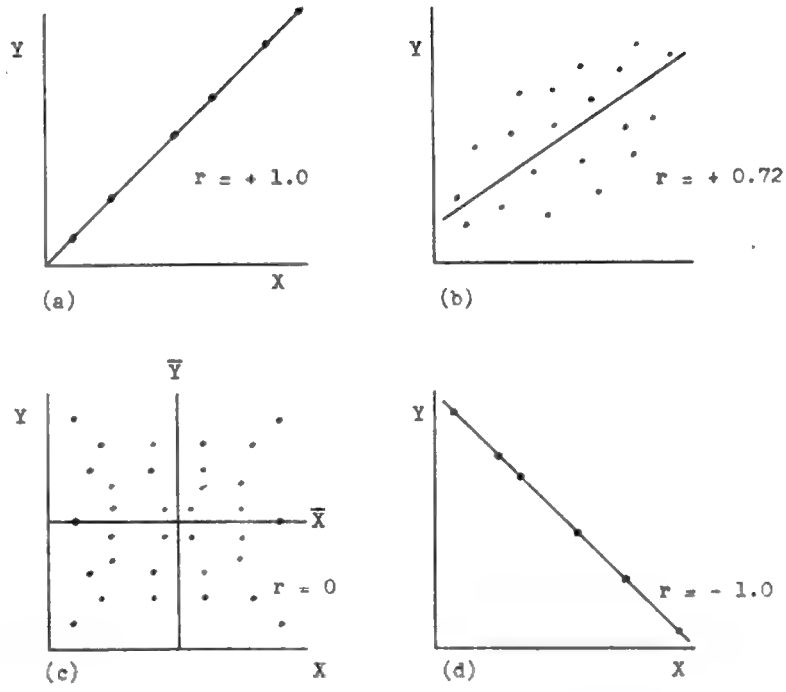


Figure 8.12 - Examples of coefficients of correlation ranging from r = 0 to r = ± 1.0.

In Figure 8.12(a), all points lie on the regression line and there is perfect correlation between Y and X; r = + 1.0. In Figure 8.12(b), the points lie in a broadened ellipse and the correlation is less than perfect, r = + 0.72. Figure 8.12(c) shows no correlation whatsoever; r = 0 and a line horizontally at $\bar{Y}$ or vertically at $\bar{X}$ would do equally as well to describe the relationship. Figure 8.12(d) shows a negative (downward to the right) perfect correlation.

Correlations of $r \geq \pm 0.6$ are fairly easy to detect; the direction of the regression line can be seen and the spread of the points in the ellipse gives an indication of the degree of correlation. As the correlation becomes less, the ellipse broadens to the point where, at $r \leq \pm 0.3$, it is difficult to tell whether the regression line should be positive or negative.

With this short introduction to correlation, let us look at r more closely. It has the formula: -

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}} \qquad (8.16)$$

where r = coefficient of correlation

Σxy = sum of the cross products $\Sigma (X - \bar{X})(Y - \bar{Y})$

Σx^2 = sum of squared deviations $\Sigma (X - \bar{X})^2$

Σy^2 = sum of squared deviations $\Sigma (Y - \bar{Y})^2$

If we examine the formula more closely, we can see that it is made up of some elements of linear regression. Previously, we had seen that the regression coefficient of Y on X was $b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$; similarly, we could say that the regression of X on Y would be $b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$.

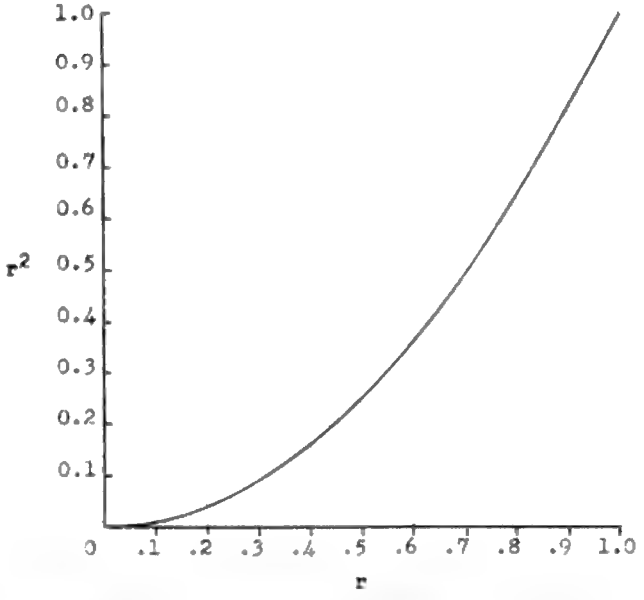


Figure 8.13 - Relation between r and r^2.

Now, let us expand r as follows: -

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

square each side:

$$r^2 = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y}} \cdot \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{\Sigma xy}{\Sigma x^2} \cdot \frac{\Sigma xy}{\Sigma y^2}$$

$$= b_{yx} \cdot b_{xy} \qquad (8.17)$$

Thus, the square of r is equal to the product of the regression coefficients b_{yx} and b_{xy}.

An important characteristic of r^2 is that it indicates the proportion of the sum of squares which are explained by regression. The relation between r and r^2 is not linear, however, as is seen by Figure 8.13.

If r = 0.4, it means that only 16% of the variation in Y is explained by the regression of Y on X. This is hardly cause for much elation, because 84% of the total variation in Y is still unaccounted for. If r = 0.82, we have 67.24% of the variation in Y explained, and this is considered to be indicative of a high correlation between Y and X. Some biological phenomena show r = 0.9 or better but these instances are rare. Nash (1963) showed that site index of shortleaf pine in Oregon County, Mo. could be estimated by using a single independent variable of a combination of slope and aspect in which the regression equation was $\hat{Y} = 29.76 + 0.979X_1$, with r = 0.580.

Test of significance of r

A test of significance of the coefficient of correlation r is given by the formula: -

$$t = \frac{r}{\sqrt{1 - r^2/n - 2}} \qquad (8.18)$$

with d.f. = n - 2. We are testing the hypothesis that $\rho = 0$ where ρ is the population coefficient of correlation. If the value of the calculated 't' is greater than the tabulated 't' at the 5% level, the coefficient of correlation is said to be significant at the 5% level. In some texts, for instance Steel and Torrie (1960), you will find a table showing the actual values of r which are significant at the 5% and 1% levels for various degrees of freedom and for different numbers of independent variables. If we use two or more independent variables, we are concerned with multiple regression and multiple correlation which will not be covered in this text.

Significance and meaningfulness

The absolute value of r at the 5% level decreases as the degrees of freedom increase; in other words, a lower value of r is significant when d.f. = 40 than if d.f. = 10. This fact requires caution when interpreting r. A significant value of r when the degrees of freedom is large can be actually meaningless as far as a practical application is concerned. Take the case where d.f. = 98; for a single independent variable, r is significant at the 5% level when it has a value of 0.195. This means that $(0.195)^2$ or 3.8% of the variation in Y has been accounted for, which will not be cause for confidence in estimating $\hat{Y}$. On the other hand, if

d.f. = 4, an r of 0.811 is significant at the 5% level, indicating that approximately 65% of the variation in Y has been accounted for.

If r^2 is the percent of variation in Y which can be accounted for by the regression of Y on X, then the percent which is not accounted for must be due to the cause or causes which we have not measured. This latter amount is given a special term called index of alienation or index of non-correlation and is indicated by $1 - r^2$.

Example of linear correlation

In order to present an example of computing the coefficient of linear correlation, we should compute the linear regression equation at the same time. The following table gives measurements of X and Y, together with columns necessary for the computation of both the regression equation and the coefficient of correlation.

Table 8.12 Data necessary for computing the linear regression equation and linear coefficient of correlation.

	X	Y	x	x^2	y	y^2	xy
	2	6	- 13	169	- 8	64	104
	3	1	- 12	144	- 13	169	156
	4	10	- 11	121	- 4	16	44
	7	4	- 8	64	- 10	100	80
	8	7	- 7	49	- 7	49	49
	9	9	- 6	36	- 5	25	30
	10	14	- 5	25	0	0	0
	11	17	- 4	16	+ 3	9	- 12
	11	15	- 4	16	+ 1	1	- 4
	12	7	- 3	9	- 7	49	21
	14	10	- 1	1	- 4	16	4
	15	6	0	0	- 8	64	0
	17	12	+ 2	4	- 2	4	- 4
	17	18	+ 2	4	+ 4	16	8
	18	13	+ 3	9	- 1	1	- 3
	18	25	+ 3	9	+ 11	121	33
	23	24	+ 8	64	+ 10	100	80
	24	20	+ 9	81	+ 6	36	54
	24	23	+ 9	81	+ 9	81	81
	27	22	+ 12	144	+ 8	64	96
	28	18	+ 13	169	+ 4	16	52
	28	27	+ 13	169	+ 13	169	169
Sums:	330	308	0	1384	0	1170	1038
Means:	15	14					

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{1038}{1384}$$

$$= 0.7500$$

$$(\hat{Y} - \bar{Y}) = b(X - \bar{X})$$

$$\hat{Y} - 14 = 0.7500(X - 15)$$

$$= 0.7500X - 11.2500$$

$$\hat{Y} = 2.7500 + 0.7500X$$

(Continued on next page)

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{1038}{\sqrt{(1384)(1170)}}$$

$$= \frac{1038}{\sqrt{1619280}}$$

$$= \frac{1038}{1273}$$

$$= 0.8154$$

$$t = \frac{r}{\sqrt{1 - r^2/n - 2}}$$

$$= \frac{0.8154}{\sqrt{\frac{1 - 0.665}{20}}}$$

$$= \frac{0.8154}{\sqrt{0.01675}}$$

$$= \frac{0.8154}{0.1294}$$

$$= 6.30 \text{ **}$$

As a result of the computations, we conclude that the coefficient of correlation is significant at the 5% level; actually, it is significant at the 1% level. The hypothesis of zero correlation is not accepted and we can state that the two variables are significantly correlated. The plotted points and the regression line are shown in Figure 8.14.

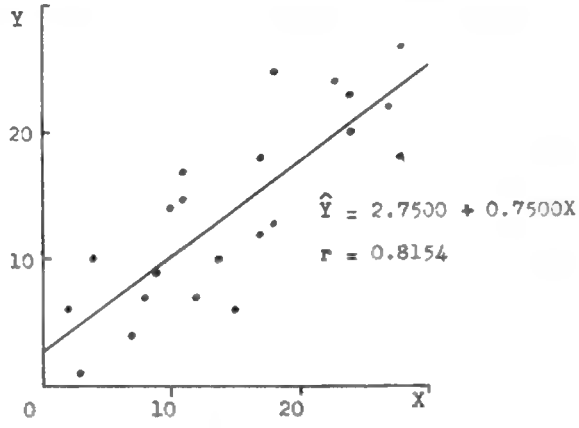


Figure 8.14 - Regression equation $\hat{Y} = 2.7500 + 0.7500X$ and $r = 0.8154$.

Chapter 9

INTRODUCTION TO EXPERIMENTAL DESIGN AND ANALYSIS OF VARIANCE

Principles of experimental design

When we speak of experimental design, we refer to the manner in which an experiment is carried out, to the number of sampling units to be used, to the precision of the results. It might be well to define an "experiment" at the beginning of this chapter so that it is clear what is meant by the term.

Experiment - an investigation which is planned to obtain new facts or to confirm or deny the results of a previous investigation.

Most experiments have the goal of learning new facts which will ultimately be of benefit to mankind.

Type of experiments

Experiments can be conveniently grouped into three general classes: -

1. exploratory
2. critical
3. demonstrational.

The exploratory experiment is one from which the investigator hopes to obtain leads for future work. Such an experiment might be conducted to determine which of a number of chemicals is best for reducing hardwood sprouting in a softwood plantation. It is assumed that a large number of different chemicals is available but that the toxicity of each is unknown. From the exploratory experiment, there might be four chemicals which show promise; this is all the exploratory experiment is designed to determine.

It is up to a critical experiment to determine which of the four chemicals gives the best results after a carefully controlled test. The experiment must be designed so that the treatment effects (chemicals) can be evaluated with statistical assurance. Each treatment is applied to a sufficiently large number of individuals so that the effect of the treatments can be measured and compared.

After determining which of the chemicals gives the best results, a demonstrational experiment is often conducted to show the value of the chemical on a practical basis. The demonstrational experiment does not add new facts; it sometimes is used to compare the derived results of the critical experiment with a standard procedure.

Most experiments which evaluate treatment effects are critical experiments.

Objectives of experiments

Before any experiment is conducted, the population must be defined; this is an extremely important consideration. The investigator must confine his sampling to the population, otherwise the results will be badly confounded. Applying treatments to more than one population will give certain information on treatment effects, but part of the difference due to treatments may be caused by having sampled different populations rather than by the treatments themselves.

"Student" (1908) states that: -

> "...any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong."

A population should be so defined as to be entirely appropriate to the purpose of the sampling; in addition, it must be usable to the person collecting a sample. In forestry, there are usually no great problems attached to the definition of a population. Some examples are 30-year old stands of white oak, 0-1 spruce seedlings from seed taken in northern Minnesota, 14-inch pine logs in a log deck, wood fibers taken from sections of oak at dbh etc.

Experimental error

One characteristic which is common to all experimental material is that of variation. Experimental error is a measure of the variation between individuals which have been treated alike. No two individuals will react exactly alike even though they have been given the same treatment. You realize, of course, that this refers to biological material, not manufactured articles. There is an inherent variation which makes individuals behave differently. Part of this variation may be due to inherited characteristics and part to the fact that it is almost impossible to obtain or measure two or more biological units which are identical. In addition to this natural variation, lack of uniformity in the physical conduct of an experiment will result in variation. Treatments may consist of the application of a chemical; can we be absolutely sure that each sampling unit receives exactly the same amount of the chemical? Are the environmental conditions uniform for each unit? temperature? humidity? moisture? Careful experimental control will reduce variation caused by lack of uniformity but it may not control the variation completely. In order to arrive at results which are meaningful, the experimental error must be reduced to a minimum; in other words, treatment differences should be caused by the treatments, not by factors over which the investigator has no control. Experimental error does not include human error in the form of mistakes in measurement, faulty application of treatments or errors in calculations. It is assumed that these human errors can, and will be eliminated.

Replication

When a treatment appears more than once in an experiment, it is said to be replicated. For instance, the same treatment may be applied to five sampling units selected at random from a population of sampling units. The treatment is said to have five replications or to be replicated five times. Do not think of a replication as a duplication of effort; it is not. There are important reasons for replication in an experiment. They are: -

1. to provide an estimate of experimental error
2. to improve the precision in estimating treatment effects.
3. to afford control of error variance.

If there is a single replicate - a treatment applied to one sampling unit only - there can be no estimate of experimental error. Differences between treatments with single replication may be due to the treatment effect or it may be due to the inherent variation in the sampling units; it is not possible to separate the two. We must be careful in increasing the replications however; as the number of replications increases, we might inadvertently be sampling a population with wider limits than was originally specified, with the results that the replications might be from a new population having larger experimental error.

Methods of reducing experimental error

Since it is important to reduce experimental error to a minimum in order to get the full effect of treatments, we can look at some methods of accomplishing this. Again, we are assuming that the treatments have been applied correctly and the measurements have been obtained

without error. There are three general methods of reducing experimental error: -

1. defining the population so that it is restricted to sampling units in which the variation is small. This point refers back to the necessity of defining the population precisely. As population boundaries are reduced, experimental error will be reduced.

2. stratification is the next method of reducing experimental error. By stratification, a population with wide limits is broken up into smaller sub-populations, each of which has smaller variation than the original. Stratification is used extensively in forestry work such as sampling for inventory. The system of forest typing is nothing more than stratifying the forest into more homogeneous sub-populations. Stratification does not discard the basic principle of random samples; it merely takes into account the investigator's knowledge that a sub-population of sampling units having little variation will result in a lower value for experimental error.

3. the third method is that used in linear and curvilinear regression where two variables are related to each other in some mathematical sense. If we can calculate a regression equation for two variables, we can estimate the value of the dependent variable with greater accuracy and efficiency (reduction in experimental error) than if we did not know or could not calculate the relationship. In the chapter on linear and curvilinear regression, we were, in effect, reducing experimental error by computing the regression equations and estimating the dependent variable through the regression equation. You will recall that the sum of squares from regression, which was then developed into standard deviation from regression, was a lower figure numerically than if we had computed the sum of squares using the mean value of the dependent variable. We were reducing experimental error without actually calling it by that name.

One final thought in this introduction to experimental design is that a person might design a very elaborate experiment and not be able to carry it out because of lack of equipment or funds. The design must be tailored to the funds available and to the resources at the disposal of the experimenter. If funds and equipment are the controlling factors in a design, the experimenter has two choices, either to postpone the experiment until funds are available or to reduce the number of treatments or replications.

Analysis of variance

Analysis of variance is just what it says; it is a partitioning of the variance (standard deviation squared) of a measured attribute into its component parts. It is a more powerful tool than the 't' test which is usually made on two groups to determine whether or not they are from the same population. In terms of the null hypothesis, the 't' test is a test of H_0: $\mu_1 = \mu_2$. Analysis of variance, on the other hand, can be extended to two or more groups and can partition the variance attributable to a number of effects in a very complex statistical experiment.

One-way classification

The simplest, most straight-forward type of experiment for analysis of variance is t.e one-way classification in which there is but one criterion for classification and in which there is an equal number of replicates in each treatment. The term "treatment" need not be confined to an external, applied variation but may also be regarded as one of natural classification such as physiographic factors, soil factors, temperature, humidity etc. Examples of one-way classifications in forestry are (1) different amount of fertilizer application in nursery beds (2) germination percent of seed subjected to various stratification temperatures (3) the effect of log diameter on sawing time in a mill and (4) different soil types and the ability to promote seedling height growth.

Analysis of variance is a useful tool when a completely random assignment of treatments is given to samples which come from a relatively homogenous population. The treatment effects, in this case, will likely prove to be more noticeable than if randomization were not carried out.

If all experimental units in a population were exactly alike, the effect of different treatments would be direct. This is not the case when dealing with biological populations. There is an inherent variability which results in varying response of measured variables even though they are treated alike. Thus, for a given treatment, there will be a variation in the measured attribute; part of it is due to the treatment effect and part to the inherent variability of the experimental units themselves. The first is called the treatment effect and the latter, experimental error. Experimental error is also called the residual sum of squares, discrepance or within groups sum of squares. You will notice that the term "sum of squares" is used to describe experimental error; the same is used in connection with the total variance and with treatment effects. One property of sums of squares is invaluable; it is that the sums of squares are additive, thus: -

$$S.S._{total} = S.S._{treatments} + S.S._{error}$$

General model for one-way classification

In a one-way classification, we have a number of treatments applied to a number of replicates in each treatment. The simplest form is that of completely randomized allocation of individuals to the treatments and an equal number of replicates in each. Also, we assume that the variance in each group is homogeneous.

With these assumptions, we can now prepare a general model for the analysis of variance; the treatments are designated as T_1, T_2.....T_j and the replications as R_1, R_2, R_3.....R_k. Thus the fourth replication in the second treatment would be identified as X_{24}. With this background information, the complete table for the 5 replications of 4 treatments is: -

	Treatments			
Replications	T_1	T_2	T_3	T_4
1	X_{11}	X_{21}	X_{31}	X_{41}
2	X_{12}	X_{22}	X_{32}	X_{42}
3	X_{13}	X_{23}	X_{33}	X_{43}
4	X_{14}	X_{24}	X_{34}	X_{44}
5	X_{15}	X_{25}	X_{35}	X_{45}

The total sum of squares for the entire experiment consisting of 20 individuals is: -

$$S.S._{total} = \Sigma X^2 - \frac{(\Sigma X)^2}{jk}$$

in which $\Sigma X^2 = X_{11}^2 + X_{12}^2 + X_{13}^2 + \ldots\ldots + X_{45}^2$

$(\Sigma X)^2 = (X_{11} + X_{12} + X_{13} + \ldots\ldots + X_{45})^2$

j = number of treatments

k = number of replications

The sum of squares for treatments is symbolized as: -

$$S.S._{treatments} = \frac{\Sigma X_{T1}^2 + \Sigma X_{T2}^2 + \Sigma X_{T3}^2 + \Sigma X_{T4}^2}{k} - \frac{(\Sigma X)^2}{jk}$$

In the sum of squares for treatments, we are only concerned with the four values which make up the sum of X's in each treatment. The total for each treatment is made up of five replications, so we divide by 5.

The sum of squares for experimental error, may be obtained by subtraction as shown previously or from: -

$$\begin{aligned} s.s._{error} = {} & \Sigma (X_{11}^2 + X_{12}^2 + X_{13}^2 + X_{14}^2 + X_{15}^2) - \frac{(\Sigma X_{T1})^2}{k} + \\ & \Sigma (X_{21}^2 + X_{22}^2 + X_{23}^2 + X_{24}^2 + X_{25}^2) - \frac{(\Sigma X_{T2})^2}{k} + \\ & \Sigma (X_{31}^2 \quad X_{32}^2 \quad X_{33}^2 \quad X_{34}^2 \quad X_{35}^2) - \frac{(\Sigma X_{T3})^2}{k} + \\ & \Sigma (X_{41}^2 \quad X_{42}^2 \quad X_{43}^2 \quad X_{44}^2 \quad X_{45}^2) - \frac{(\Sigma X_{T4})^2}{k} \end{aligned}$$

We now have the necessary model to follow in a one-way classification of analysis of variance and we are ready for an example.

An experiment was run in which 5 replications of 50 black oak acorns were planted in each of four soil types. At the end of the first growing season, the heights of the seedlings were measured as an indication of the effect of the soil types to promote height growth. As far as possible, all other growing conditions were kept uniform for each treatment (soil type). In order to simplify the calculations, the average height growth for each replications has been used as a single entry and the data have been altered slightly. Table 9.1 gives the data.

Table 9.1 - Data on height growth of black oak seedlings on four soil types.

	Soil type				Entire experiment
	Menfro	Weldon	Union	Marshall	
		inches			
	1.8	2.6	2.4	2.4	
	1.6	3.2	2.2	2.6	
	2.1	3.3	2.3	2.1	
	1.5	3.6	2.1	2.3	
	2.0	2.8	2.0	3.1	
Sums:	9.0	15.5	11.0	12.5	48.0
Means:	1.8	3.1	2.2	2.5	2.4
ΣX^2:	16.46	48.69	24.30	31.83	121.28
$\frac{(\Sigma X)^2}{k}$:	16.20	48.05	24.20	31.25	$\frac{(\Sigma X)^2}{jk}$ = 115.20
Σx^2:	0.26	0.64	0.10	0.58	6.08

In Table 9.1, we have four treatments with five replications in each. We can visualize the experiment as either(1) one in which treatments are disregarded and all 20 measurements comprise the experiment with $\Sigma X = 48.0$ and $\bar{X} = 2.4$ or (2) one in which the effect of the treatments can be analyzed for the reduction in the sum of squares due to treatments. Each of these views will provide an estimate of variance and it is our job to examine the difference between the two estimates and to determine whether the difference between treatment means is significant. For this, we postulate the hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, and, through analysis of variance, accept the null hypothesis or reject it and set up an alternate hypothesis H_A: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$. To begin with, we shall see whether we accept or reject the null hypothesis; later, we shall carry the computations a step further if we reject H_0.

Following the procedures for obtaining sums of squares outlined on page 92, we have: -

Total sum of squares

$$S.S._{total} = 1.8^2 + 1.6^2 + 2.1^2 + \ldots\ldots + 2.3^2 + 3.1^2 - \frac{(48.0)^2}{20}$$

$$= 121.28 - 115.20$$

$$= 6.08$$

Treatment sum of squares

$$S.S._{treatments} = \frac{9.0^2 + 15.5^2 + 11.0^2 + 12.5^2}{5} - \frac{(48.0)^2}{20}$$

$$= \frac{81.00 + 240.25 + 121.00 + 156.25}{5} - 115.20$$

$$= 119.70 - 115.20$$

$$= 4.50$$

Error sum of squares

$$S.S._{error} = S.S._{total} - S.S._{treatments}$$

$$= 6.08 - 4.50$$

$$= 1.58$$

The error sum of squares can also be found by the following procedure: -

Treatment 1, Menfro soil

$$1.8^2 + 1.6^2 + 2.1^2 + 1.5^2 + 2.0^2 - \frac{9.0^2}{5}$$

$$= 16.46 - 16.20$$

$$= 0.26 \qquad 0.26$$

Treatment 2, Weldon soil

$$2.6^2 + 3.2^2 + 3.3^2 + 3.6^2 + 2.8^2 - \frac{15.5^2}{5}$$

$$= 48.69 - 48.05$$

$$= 0.64 \qquad 0.64$$

Treatment 3, Union soil

$$2.4^2 + 2.2^2 + 2.3^2 + 2.1^2 + 2.0^2 - \frac{11.0^2}{5}$$

$$= 24.30 - 24.20$$

$$= 0.10 \qquad 0.10$$

Treatment 4, Marshall soil

$$2.4^2 + 2.6^2 + 2.1^2 + 2.3^2 + 3.1^2 - \frac{12.5^2}{5}$$

$$= 31.83 - 31.25$$

$$= 0.58 \qquad 0.58$$

$S.S._{error}$: 1.58

The error sum of squares is the sum of squares remaining after the treatment effect has been taken into account. It is the inability of the treatments to take care of all the variation in the measured variable. As the value of the error term gets smaller, it means that the treatments are accounting for more of the total variation; this is precisely what we are attempting to do in an experiment by assigning treatments. If the treatment sum of squares did not result in much reduction of the total sum of squares, we would be in a position of saying that the treatments were not very effective. Statistically speaking, we would accept the null hypothesis and say that the treatments did not indicate that the sample means were from different populations.

The analysis of variance is normally contained in a table showing (1) the source of variation (2) the degrees of freedom associated with each source (3) the sum of squares for each and the (4) mean square. From these data, we are able to extract the mean squares for treatments and for the error term which are necessary to establish the ratio F.

Table 9.2 - Analysis of variance for one-way classification.

Source of variation	d.f.	Sum of squares	Mean square
Total	19	6.08	
Treatments	3	4.50	1.50
Error	16	1.58	0.0987

$$F_{3,16} = \frac{1.50}{0.0987} = 15.2\ **$$

Looking at the F-table (Table A.4 in Appendix) for d.f. 3 and 16, we find the 5% level of probability is 3.24 and for the 1% level, 5.29. Our calculated value of 15.2 is greater than either of these, so we conclude that there is less than 1 chance in 100 that a greater value of F would occur by chance. On this basis, we reject the null hypothesis and accept the alternate hypothesis that $H_A: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$. This is the only hypothesis we can formulate at the present time. All we know is that there is a significant difference (at the 1% level) between treatment means but we do not have any idea which treatment or treatments is responsible for our rejecting the null hypothesis. It is important to the success of an experiment to isolate the treatment or treatments which cause us to reject a null hypothesis.

A method of analyzing the difference between the means of the four treatments has been supplied by Snedecor (1956) and is called the Q-test.

Snedecor's Q-test

A difference, D, is computed and then compared to a combination of differences between the sample means. The D is the result of standard error and a factor Q, such that: -

$$D = Q\, s_{\bar{x}} \qquad (9.1)$$

The $s_{\bar{x}}$ can be easily computed from $s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$ in which s^2 is the mean square of the error term and n is the number of replications. In the height growth example, s^2 is 0.0987 and n is 5, giving $s_{\bar{x}}$ the value of $\sqrt{\frac{0.0987}{5}} = 0.1406$. The value of Q is obtained from Table 9.3 which shows the upper 5% of the range for different degrees of freedom and for number of

treatments. In order to determine which treatment or treatments is causing the significant difference between means, we use a different value of Q depending on whether the treatment means are 2, 3, or 4 ranks apart. In Table 9.3, for d.f. = 16, use Q = 3.00 for adjacent means (a = 2), Q = 3.65 for means three ranks apart (a = 3) and Q = 4.05 for means four ranks apart (a = 4).

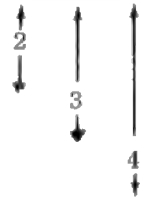
Table 9.3 - Upper 5% points, Q.

Degrees of freedom	Number of treatments, a								
	2	3	4	5	6	7	8	9	10
1	18.0	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.08	8.33	9.80	10.88	11.74	12.44	13.03	13.54	13.99
3	4.50	5.91	6.82	7.50	8.04	8 48	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.44	4.76	5.02	5.24	5.43	5.59	5.74
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47

Table 9.3 is taken from J. Parchares "Table of the Upper 10% Points of the Studentized Range", Biometrika 46: 461-466 (1959) with additional table by Dr. Leon Harter by permission of Professor E. S. Pearson, Editor of Biometrika.

When the means are ranked in order from highest to lowest, we have: -

Order	Soil	Mean	Ranks apart
1	Weldon	3.1	
2	Marshall	2.5	
3	Union	2.2	
4	Menfro	1.8	



Determine $D = Qs_{\bar{x}}$ for each value of Q.

$$a = 2 \quad D = 3.00(0.1406) = 0.422$$

$$a = 3 \quad D = 3.65(0.1406) = 0.513$$

$$a = 4 \quad D = 4.05(0.1406) = 0.569$$

Now rank the means in order and determine the differences between the highest and lowest.

Soil type	Mean	$\bar{X}$ - 1.8	$\bar{X}$ - 2.2	$\bar{X}$ - 2.5
Weldon	3.1	1.3 (0.569)	0.9 (0.513)	0.6 (0.422)
Marshall	2.5	0.7 (0.513)	0.3 (0.422)	
Union	2.2	0.4 (0.422)		
Menfro	1.8			

For simplicity and to demonstrate the differences between means and the comparisons with D, the D values have been inserted in parentheses. If the difference between means is greater than the value of D for a particular difference, the difference between means is significant at the 5% level. For instance, $\bar{X}$ - 1.8 = 1.3 is greater than D = 0.569, therefore the difference between the means of the Weldon and Menfro soils is significant, but the difference between the means of the Marshall and Union soils is not. The complete list of differences is as follows: -

Between	and	Difference is
Weldon	Menfro	significant
Weldon	Union	significant
Weldon	Marshall	significant
Marshall	Menfro	significant

Our conclusion is that the Weldon soil produces significantly different height growth from all other soils tested and that the Marshall soil is significantly different from the Menfro soil. There is no significant difference between Union and Menfro or between the Marshall and Union. This information is more than we obtained from a straight analysis of variance, which was that a significant difference occurred due to treatment (soil type). Now we know that the Weldon soil provided the greatest difference and this fact alone could be instrumental in administrative decisions on the future planting of black oak.

Two-way classification

In two-way classification, there are two treatments to take into account; each of them has some effect on the values of the measured variable. It is our task to determine whether the treatments have a significant effect on the outcome. As in the single classification, analysis of variance is the tool to use.

The most direct case of a two-way classification problem is that called a randomized block design. In addition to a treatment effect caused by different levels of a treatment, the experimental material may be grown in different localities or under different environmental conditions. It is obvious that some differences in reaction might take place if the same treatment were applied under varying growing conditions. You might ask the following question - why not make sure that the external conditions are uniform for all treatments? Even in a highly controlled situation like a greenhouse or in a laboratory, it is almost impossible to have all conditions uniform. Remember that we are trying to reduce the experimental error term to a minimum; the greater is the sum of squares attributable to known causes, the smaller is the error sum of squares. Since the test of significance is the F-test in which the mean square from a known cause is divided by the mean square for the error term, it stands to reason that the lower we can make the error mean square, the higher is the probability of detecting a significant effect for treatments or conditions.

In a randomized block experiment, treatments are assigned to blocks; within each block, there are plots to which are assigned the treatments. There are as many plots within blocks as there are treatments. The variation between blocks will form a part of the experimental variation and there always exists the possibility of some

	Block 1	Block 2	Block 3	Block 4	Block 5
Plot 1	A	B	C	E	D
Plot 2	C	D	E	B	C
Plot 3	D	A	B	C	E
Plot 4	B	E	A	D	A
Plot 5	E	C	D	A	B

Figure 9.1 - Random assignment of five treatments to plots within five blocks.

differences between blocks which will cause a differential response in the measured variable. Consider a case where we have five treatments; it does not matter at the present time, what the treatments are, so let us call them A, B, C, D and E. Now set up five blocks, each with five plots and assign the treatments at random to the plots within the blocks. It is not necessary to have the same number of blocks as there are treatments, only that the number of plots within blocks be the same. We may have the following situation as shown in Figure 9.1.

If position within each block were a source of variation, the random assignment of treatments to plots should eliminate it.

In a randomized block experiment, we have four possible sources of variation, each producing its own estimate of variance. They are: -

1. total sum of squares
2. treatment sum of squares
3. block sum of squares
4. experimental error sum of squares.

This design represents an increase in efficiency because we are removing from the total sum of squares one additional source of variation, that of blocks.

Example of a randomized block experiment

The situation for the experiment is as follows: -

> Shortleaf pine seed is available from four sources, Arkansas, Kentucky, Tennessee and Texas. The seedling survival for each is desired under Missouri conditions.

The procedure for the physical handling of the experiment would be to select a random sample from each seed source and assign each sample at random to a plot within each of five blocks. The blocks may be beds in a forest nursery or flats in a greenhouse which may or may not be adjacent to each other. The random assignment can be represented as shown in Figure 9.2.

	Block 1	Block 2	Block 3	Block 4	Block 5
Plot 1	Ark. 86	Ky. 81	Texas 70	Tenn. 70	Ky. 71
Plot 2	Tenn. 74	Texas 68	Ky. 74	Ark. 82	Tenn. 71
Plot 3	Ky. 76	Ark. 83	Tenn. 73	Texas 65	Texas 66
Plot 4	Texas 71	Tenn. 72	Ark. 78	Ky. 73	Ark. 76

Figure 9.2 - Random assignment of four seed sources to plots within five blocks. Survival percentages shown in each plot.

The survival, which is to be tested, is then determined after a predetermined length of time and expressed as a percentage. The data are given in Table 9.4.

Table 9.4 - Survival percentage by seed source in plots within blocks

	Seed source				
Block	Arkansas	Kentucky	Tennessee	Texas	Sum
			Percent		
1	86	76	74	71	307
2	83	81	72	68	304
4	78	74	73	70	295
4	82	73	70	65	290
5	76	71	71	66	284
Sum:	405	375	360	340	1480
Mean:	81	75	72	68	74

Examine Table 9.4; there is a drop in the average percent survival from left to right with the greatest difference between the averages being 13%. Also, there is a decrease in survival percent as we progress through Blocks 1 to 5. At the present time, we do not know whether these differences are significant or not and we must examine the date by analysis of variance.

As stated previously, we require a sum of squares for all sources of variation, starting with: -

1. Total s.s.

$$86^2 + 83^2 + 78^2 + \ldots\ldots + 65^2 + 66^2 - \frac{1480^2}{20}$$

$$= 110128 - 109520 = 608$$

2. Treatment s.s.

$$\frac{405^2 + 375^2 + 360^2 + 340^2}{5} - \frac{1480^2}{20}$$

$$= 109970 - 109520 = 450$$

Note: Each entry is composed of the addition of five block measurements, so each is divided by five to obtain the average.

3. Block s.s.

$$\frac{307^2 + 304^2 + 295^2 + 290^2 + 284^2}{4} - \frac{1480^2}{20}$$

$$= 109611.5 - 109520 = 91.5$$

Note: Each entry is composed of the addition of values for four treatments, so is divided by 4 to obtain the average value.

4. Experimental error s.s.

Experimental error sum of squares is obtained by subtraction.

Error s.s. = Total s.s. - (Treatment s.s. + Block s.s.)

$$608 - (450 + 91.5)$$

$$608 - 541.5 = 66.5$$

These sums of squares are then tested for significance by: -

1. obtaining the mean square for each

$$\frac{\text{Sum of squares}}{\text{d.f.}}$$

2. applying the F-test on each mean square

$$F = \frac{\text{Mean square for source of variation}}{\text{Error mean square}}$$

The analysis of variance is shown in Table 9.5.

Table 9.5 - Analysis of variance for randomized block design using four treatments in five blocks.

Source of variation	d.f.	Sum of squares	Mean square	F
Seed source	3	450	150	27.1 **
Blocks	4	91.5	22.8	4.1 *
Error	12	66.5	5.6	
Total	19	608		

The F for treatments is computed from: -

$$F_{3,12} = \frac{\text{Treatment mean square}}{\text{Error mean square}}$$

$$= \frac{150}{5.6}$$

$$= 27.1$$

Looking in the F-table under d.f. = 3 and 12, we find that the 5% level of significance is given as 3.4 and the 1% as 5.7. Our calculated value of F - 27.1 - is considerably larger than the tabulated 1% value, so we conclude that seed source did cause a significant effect on survival percentage. The F-value for blocks is significant at the 5% level, but not at the 1%, indicating that there was a difference in performance between blocks. Apparently, the physical location of the blocks or the environmental conditions between the blocks were sufficiently different to cause a significant difference in survival percentage.

In the randomized block experiment, we have removed the sum of squares attributable to environment (blocks) and so have increased the efficiency of the design. This is a decided advantage over the pure random design used in the previous example. The one possible disadvantage to the randomized block design is that it gets rather unwieldy when there is a large number of treatments.

You may have noticed some similarity between this experiment and the one presented for the one-way classification, especially in the sum of squares for total and treatments. This present design is a "cook book" type which used essentially the same data except that it was coded so that $X_{2\text{-way}} = 10X_{1\text{-way}} + 50$. Let us compare the sums of squares for both experiments.

	Sum of squares	
Source of variation	One-way classification	Randomized block design
Treatments	4.50	450
Blocks	-	91.5
Error	1.58	66.5
Total	6.08	608

The randomized block design removed $\frac{91.5}{158}$ or 58% of the original error sum of squares, a decided improvement in the experimental design.

This completes the discussion of the simplest form of the analysis of variance for the two-way classification and we are now ready to examine a more complicated form called the factorial type of experiment.

Factorial experiments

When two or more treatments, each with various levels, are applied to experimental material, the experimental design is called a factorial type. The treatments can be thought of as factors; thus, a 2-factor experiment with 3 levels of each is called a 3x3 factorial. Other examples of this notation follow. The factors are conveniently called A, B, C etc. and the levels, 1, 2, 3 etc. Thus, a 2-factor experiment with 3 levels in each has the following combinations, A_1B_1, A_1B_2, A_1B_3, A_2B_1, A_3B_1, A_2B_2, A_3B_2, A_2B_3 and A_3B_3.

Number of factors	Levels	Notation
2	2 in each	2x2 or 2^2
3	2 in each	2x2x2 or 3^2
4	3 in each	3x3x3x3 or 4^3
2	2 in A, 4 in B	2x4
3	3 in A, 2 in B, 3 in C	3x2x3

The factorial design permits analysis of variance, not only for the main factors, but also for the combination of the factors. The term interaction has been given to the effect of the combination of factors and is an additional reduction in the total sum of squares.

We would like the treatments (factors) to work independently of each other so that the effect of one treatment would be the same for the various levels of the other in a two-way experiment. This does not always happen, however; if there is a lack of independency, the factors interact and there is a differential effect. The analysis of variance will determine whether the interaction term is significant or not. If it is not, it means that the factors are leaning toward an independent status and that the effect of one factor would produce approximately the same results on the measured variable at each level of the other factor. The effect of interaction will be brought out in the example demonstrating a 2x2 factorial.

A factorial experiment consists of measurements on the experimental data applied to plots within blocks. The number of plots is equal to the possible number of factors and levels of factor combinations. A 2x2 factorial would have four plots per block because we have the possibility of two levels of factor A with each level of factor B. The purpose of the blocks is to provide another estimate of variance in which it is possible to isolate any environmental or geographic location effect as a source of variation. If there were only one block, we could not achieve this result, and whatever effect location might have would be submerged in the error term.

Taking a 2x2 factorial as an example of this type with two levels of each, we have: -

	Level	
Factor A	1	2
Factor B	1	2

so the possible combination of the factors is: -

$$A_1B_1,\quad A_1B_2,\quad A_2B_1,\quad A_2B_2.$$

Each of these combinations would then be applied at random to one of the four plots in as many blocks as might be deemed necessary. The number of blocks is not fixed, and may be controlled by physical condition of the experiment, by financial support for the project, or a combination of both.

A typical 2x2 factorial with three blocks would appear as shown in Figure 9.3.

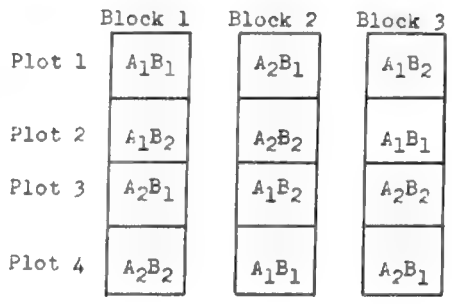


Figure 9.3 - Random allocation of 2 treatments at 2 levels each in 3 blocks.

Estimates of variance in a 2x2 factorial is obtained from: -

1. total sum of squares
2. treatment (factor) sum of squares, made up of: -
   a. treatment A sum of squares
   b. treatment B sum of squares
   c. interaction sum of squares
3. block sum of squares
4. error sum of squares.

We have made another advance in reducing the total sum of squares because each factor and the interaction between the factors are now contributing their own sums of squares. Remember that our objective is to reduce the sum of squares for the error term as much as possible. With each additional sum of squares we can remove from the total, we are increasing the possibility of detecting significant effects.

Example of a 2x2 factorial

Following the same approach as the randomized block experiment, let us take one seed source and subject the seeds to two treatments at two levels each. The treatments need not be applied in the sense of a physical addition of a chemical but may be conditions such as different soil types, various lengths of artificial light or different amounts of shade.

Using the Arkansas seed source, the survival of seedlings is to be tested for two soil types (Weldon and Menfro) and two soil conditions (woods and old field) on three blocks. The combination of factors will follow that shown in Figure 9.3. After a designated time lapse, the seedling survival is computed for each plot within the blocks and expressed as a percentage. Table 9.6 gives the data at the conclusion of the field experiment.

Table 9.6 - Data for a 2x2 factorial showing percent survival on two levels of soil type and two levels of soil condition for three blocks.

Soil type (A)	Soil condition (B)						
	Woods			Old field			Sum
	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3	
Weldon	82	86	80	74	80	76	478
Menfro	75	74	68	70	62	65	414
Sum by blocks	157	160	148	144	142	141	
Sum by conditions	465			427			892

It is helpful to make two additional breakdowns of the original table to use in later computations. The first is by treatment. Ignore the block entries and make a table showing the combined results of the three blocks under each combination of treatments. Table 9.7 gives the data; the first entry is composed of 82 + 86 + 80 = 248, and on.

Table 9.7 - Data from Table 9.6 showing results by treatments.

	Soil condition (B)		
Soil type (A)	Woods	Old field	Sum
Weldon	248	230	478
Menfro	217	197	414
Sum	465	427	892

The second is to show block totals for the two levels of soil type. Take each block within a soil type and add the entries; Block 1 is then made up of an entry for Soil type 1, Soil condition 1 and 2; Block 2 has an entry for Soil type 1, Soil condition 1 and 2 and so on. The first entry in Table 9.8 is from (Weldon-Woods Block 1) + (Weldon-Old field Block 1). Thus, the entry 156 is made up of 82 + 74.

Table 9.8 - Data from Table 9.6 showing results for blocks by soil type.

Soil type	Block 1	Block 2	Block 3	Sum
Weldon	156	166	156	478
Menfro	145	136	133	414
Sum	301	302	289	892

The sums of squares necessary for the analysis of variance are:

1. Total s.s.

$$82^2 + 75^2 + 86^2 + \ldots\ldots + 76^2 + 65^2 - \frac{892^2}{12}$$

$$= 66866 - 66305 = 561$$

2. Treatment s.s.

$$\frac{248^2 + 230^2 + 217^2 + 197^2}{3} - \frac{892^2}{12}$$

$$= 66760 - 66305 = 455$$

3. Block s.s.

$$\frac{301^2 + 302^2 + 289^2}{4} - \frac{892^2}{12}$$

$$= 66330 - 66305 = 25$$

4. Treatment A s.s. (soil type)

$$\frac{478^2 + 414^2}{6} - \frac{892^2}{12}$$

$$= 66640 - 66305 = 135$$

5. Treatment B s.s. (soil condition)

$$\frac{465^2 + 427^2}{6} - \frac{892^2}{12}$$

$$= 66420 - 66305 = 115$$

6. Interaction s.s.

The interaction sum of squares can be obtained by subtraction. The treatments sum of squares (455) is composed of three parts: -

a. Treatment A s.s.

b. Treatment B s.s.

c. Interaction s.s.

We have calculated two out of these three and we know the value of the Treatments s.s., so: -

Interaction s.s. = Treatments s.s. - (Treatment A s.s. + Treatment B s.s.)

= 455 - (135 + 115)

= 205

7. Error s.s.

The error s.s. is the remainder after subtracting (1) Treatments s.s. and (2) Block s.s. from the Total s.s. Thus: -

Error s.s. = Total s.s. - (Treatments s.s. + Block s.s.)

= 561 - (455 + 25)

= 81

Degrees of freedom

In order to determine the mean square for each component contributing to variance, it is necessary to apply the correct degrees of freedom. We know that the degrees of freedom for Total sum of squares is one less than the number of individual plots or 12 - 1 = 11. Follow Table 9.9 for the remaining degrees of freedom.

Table 9.9 - Degrees of freedom for sources of variation in a 2x2 factorial.

Source of variation		d.f.
Blocks		2
Treatments (see Note 1 below)		3
Treatment A	1	
Treatment B	1	
Interaction (See Note 2)	1	
Error		6
Total		11

Note 1: While there are only two treatments, A and B, each is at two levels, so there is a total of four possible combinations of treatments, resulting in d.f. = 3.

Note 2: The degrees of freedom for interaction is the product of the degrees of freedom for each treatment.

We now have all the necessary information to complete the analysis of variance.

Table 9.10 - Analysis of variance for a 2x2 factorial with three blocks

Source of variation		d.f.	Sum of squares	Mean square	F
Blocks		2	25	12.5	0.823
Treatments		3	455	151.6	10.05 **
Treatment A	1		135	135.0	8.70 *
Treatment B	1		115	115.0	7.60 *
Interaction AB	1		205	205.0	13.56 *
Error		6	81	15.1	
Total		11	561		

Interpretation and discussion of the results

Let us examine the results of this analysis of variance and determine what we can say about (a) the way in which the experiment was designed and (b) the meaning of the various F-values shown in Table 9.10.

Starting with the conduct of the experiment, we can examine the original data (Table 9.6) and formulate some ideas as to the effectiveness of the factors. At first glance, it seems reasonable to assume that there might be a significant difference between the two soil types by comparing the means by types, $\frac{478}{6} = 79.6$ vs. $\frac{414}{6} = 69.0$. The same conclusion might be formulated by comparing the means for soil conditions, $\frac{465}{6} = 77.5$ vs. $\frac{427}{6} = 71.2$. How about the block means? Referring to Table 9.8, we see that the block means are $\frac{301}{4} = 75.1$, $\frac{302}{4} = 75.2$ and $\frac{289}{4} = 72.2$; there is not a very large range between the highest and lowest mean. This fact alone would indicate that the block effect will not be very strong. Remember that these are but thoughts on the possible outcome and we have no evidence as yet. It appears on the surface that our choice of treatments (soil types and soil conditions) was reasonable.

Now let us examine the results, particularly the F-values. In Table 9.10, we notice that all the mean squares are significant. The mean square for Treatments is significant at the 1% level and at the 5% level for Treatment A, Treatment B and for interaction. The mean square for Blocks is not significant. Apparently, our separation of the experimental data into blocks did not cause any significant differences in the survival of the seedlings because of any environmental cause.

As for the soil type treatment (Factor A), there is a significant difference between the Weldon and Menfro soils. This is as far as we need go in this interpretation; with only two levels of a factor, an F at the 5% level implies that one is significantly different from the other. With only two levels of a factor being considered, we could have reached the same conclusion by determining the value of 't' because in this particular case, $F = t^2$.

The same reasoning can be applied to soil condition (Factor B) and we can state that the woods condition produced significantly better survival rates than the old field condition.

If we had had three levels or more of each factor, the interpretation would not be as

simple and straight-forward. An F for treatments at the 5% or 1% level would imply a significant difference between treatments, but we would have to resort to Snedecor's Q test or some other test such as Duncan's multiple range test to determine which of the treatments was causing the significant difference.

In the above interpretation, we are assuming the risk that we could be wrong in our conclusion 1 time in 100. We are actually setting up hypotheses that: -

1. Blocks: H_0: $\mu_1 = \mu_2 = \mu_3$

2. Treatment A: H_0: $\mu_{A_1} = \mu_{A_2}$

3. Treatment B: H_0: $\mu_{B_1} = \mu_{B_2}$

4. Interaction: H_0: $\mu_{A_1B_1} = \mu_{A_1B_2} = \mu_{A_2B_1} = \mu_{A_2B_2}$

As a result of the F-values, we fail to reject the null hypothesis for blocks and say that there is a lack of evidence to show that the block means came from different populations. In the case of the remaining sources of variation, we reject the null hypotheses and postulate alternate ones such as H_A: $\mu_1 \neq \mu_2$ and so on. The probability of making an error in judgment is 1%; if you look up a Type I error (page 44), you will see that it involves the rejection of a null hypothesis when it should be accepted, and a Type II error is accepting a null hypothesis when it should be rejected. In either case, the probability of making an error should be low. A level of significance of say, 0.05, states that this is the maximum probability we would be willing to risk in making an error in judgment.

This section completes the work to be covered in this text. For more detailed study of advanced statistical theory and application, the student is referred to Snedecor (1956), Steel and Torrie (1960), Edwards (1950), Cochran (1953), Cochran and Cox (1950) and Fisher (1942) for such subjects as 3x3x3 factorials, multiple regression and correlation analysis, covariance, Latin squares and other more advanced experimental designs.

APPENDIX

LABORATORY EXERCISES

These exercises are designed to supplement various parts of the text in a more formal manner than the problems. The exercises can normally be completed in a 2 hour laboratory session and should be written up to conform to standard procedures in presenting reports.

Exercise 1

OPERATION OF A DESK CALCULATOR

Electric or hand-operated desk calculators are used a great deal in forestry work and in statistics. Usually different makes and models of calculators are available for student use and it is essential that the student becomes familiar with their operation.

Perform the following calculations and record the results in the spaces provided: -

Addition

Add the following list: -

1. 303.46
2. 13.03
3. 1460.2028
4. 6.52
5. 249.837
6. 0.01
7. 63.4
8. 582.472
9. 78.0
10. 146.302

Total _____

Multiplication

1. 492.06 x 6.302 _____
2. 0.0062 x 4.01 _____
3. 100.06 x 55.11 _____
4. 65.20 x 10.16 x 0.05 _____
5. 18,530 x 0.1092 _____

Division

1. 185 ÷ 16 _____
2. 10,532 ÷ 12.66 _____
3. 538.0 ÷ 428.20 _____
4. 4.083 ÷ 8,576.29 _____
5. 986 ÷ 12,841.24 _____

Squares

Cumulate the squares for each of the following lists: -

1	2	3
53	1.41	0.7
82	3.02	1.1
49	2.67	1.6
76	1.29	2.3
69	0.62	3.9
51	9.40	4.4
46	7.81	4.9
80	6.33	5.4
42	1.02	6.2
Total		

Square roots

Use the special set of tables provided for the machine for the following square roots. Carry the answer to 3 places of decimals.

1. $\sqrt{110.465}$ _____
2. $\sqrt{1.2863}$ _____
3. $\sqrt{0.05320}$ _____
4. $\sqrt{14{,}328.549}$ _____
5. $\sqrt{628.1063}$ _____

Miscellaneous

Obtain the results of the following computations: -

1. $\dfrac{110.46 \times 8.35}{14.81}$ _____
2. $\dfrac{48.32 \times 19.72}{19.92}$ _____
3. Cumulate

   13.41 x 8.20
   16.88 x 5.46
   20.92 x 9.81
   28.12 x 9.31
   185.98 x 0.87
   101.72 x 3.46 _____

Exercise 2

DRAWING AND BALANCING A CURVE

Object:

To draw and balance a curve through a given set of points.

Materials needed:

1 sheet 10 x 10 graph paper
3H pencil
French curve

Data:

The data gives the average dbh and height for the number of trees shown in each dbh class. The data were collected from Boone County, Missouri in 1960 and represent diameters and heights on second-growth white oaks measured for Project 124b.

dbh class inches	Average dbh inches	Average height feet	Number of trees
5	4.8	29	5
6	6.3	35	4
7	7.0	39	18
8	8.2	37	25
9	8.9	46	17
10	10.1	44	14
11	10.9	41	9
12	12.2	43	6
13	13.1	51	3
14	13.9	49	2

Procedure:

1. Plot the points, using a scale which will utilize as much of the graph paper as possible. Enter the frequency of each class beside each point.

2. Draw a smooth curve so that it is approximately balanced.

3. Use the "First Estimate" table at the end of the Exercise to record the estimated heights for each dbh in Column 2.

4. Subtract the estimated height (curve value) from the actual height and record the difference in the table, retaining the algebraic sign.

5. Multiply the difference obtained in (4) by the frequency for each dbh class and record the results in the f(Diff) column.

6. Compare the positive values of f(Diff) with the negative values. If the difference between the two is greater than ± 2, the original curve must be adjusted.

7. a. If your original curve is satisfactory according to the specifications in (6), proceed to Step 8.

b. If your original curve does require adjustment, make a trial adjustment and repeat Steps 2 to 6 inclusive. Record your results in the "Second Estimate" table.

If you subtracted the estimated height from the actual height in Step 4, and a positive difference resulted from the comparison between the positive and negative values of f(Diff), it means that the curve must be raised slightly. Remember to include the class frequencies in the adjustment.

c. If necessary, adjust the curve a second time and record the results in the "Third Estimate" table.

8. Complete the graph according to the procedures given in Chapter 1 and record the heights read from the curve to the nearest 0.5 foot for each one-inch class of dbh from 5 to 14 inches inclusive.

Calculation of Effectiveness of the Curve:

Average dbh Inches	f	Average Height Feet	First Estimate			Second Estimate		
			Height	Diff	f(Diff)	Height	Diff	f(Diff)
				Sum + 's			Sum + 's	
				Sum - 's			Sum - 's	

Average dbh Inches	f	Average Height Feet	Third Estimate			Fourth Estimate		
			Height	Diff	f(Diff)	Height	Diff	f(Diff)
				Sum + 's			Sum + 's	
				Sum - 's			Sum - 's	

Exercise 3

GRAPHIC PRESENTATION OF DATA

Object:

To present tabular data in graphic form

Materials needed:

3 sheets 10 x 10 graph paper
3H pencil

Data:

The following table gives the total monthly precipitation for two locations in Missouri, Columbia and Weldon, for the year 1958.

Month	Total Monthly Precipitation[1]	
	Columbia inches	Weldon inches
January	1.7	2.1
February	1.6	1.4
March	3.9	2.4
April	2.3	1.9
May	3.4	3.8
June	4.6	4.8
July	6.4	10.4
August	2.0	2.3
September	1.8	2.9
October	2.5	0.8
November	4.1	2.7
December	0.3	0.6

Procedure:

1. Construct a bar chart to show the total monthly precipitation figures graphically.
2. Draw a frequency polygon showing the total monthly precipitation.
3. Prepare a cumulative frequency table for the Weldon data and draw the cumulative frequency polygon (ogive).

---

[1] Data extracted from Settergren, C. D. Initial survival and growth of oak seedlings in the Missouri River Hills. Unpublished thesis for Master's degree in Forestry, University of Missouri, 1959.

Exercise 4

MEASURES OF CENTRAL TENDENCY

Object:

To approximate a normal curve and to compute some measures of central tendency.

Materials needed:

1 sheet 10 x 10 graph paper
3H pencil

Data:

The following data were obtained by tossing 10 coins in the air at one time and counting the number of heads which turned up. The tosses were repeated until the total number was 750.

Number of heads turned up	Frequency
0	0
1	6
2	32
3	82
4	164
5	180
6	166
7	85
8	30
9	5
10	0
	750

Procedure:

1. Plot frequency on number of heads and draw: -
   a. the frequency polygon
   b. a smooth curve by connecting the mid-points of the classes.
2. Compute: -
   a. the mean number of heads
   b. the mode
   c. the median

   and show these values on the X-axis.
3. Compute: -
   a. Q1 and Q3 (the 1st and 3rd quartiles)
   b. D1, D3 and D7
   c. P_{22}, P_{65} and P_{92}.

Exercise 5

MEASURES OF DISPERSION

Object:

To study a frequency distribution and to determine various measures of dispersion in connection with it.

Materials needed:

Column paper
Desk calculator

Data:

Diameter at breast height of a number of trees in an even-aged stand was measured on a one-acre plot and the following distribution was recorded: -

dbh inches	Frequency
4	27
5	38
6	57
7	63
8	41
9	17
10	5
	248

Procedure:

1. Set up the appropriate columns and compute the following statistics: -

   a. the mean, using Formula 3.3

   b. standard deviation, using Formula 5.3

   c. coefficient of variation

   d. the Z value for each dbh class (Remember that $Z = \frac{X - \bar{X}}{s}$ with the deviation using the true mean, not the assumed mean).

   e. list the probability of occurrence of each dbh class by consulting Table A.2.

Exercise 6

STANDARD ERROR

Object:

To compute various statistics concerned with standard error and the use of 't'.

Materials needed:

Column paper
Desk calculator

Data:

A. The table below shows the distribution of a sample taken from a normally-distributed population: -

X	Frequency
0	2
1	8
2	16
3	25
4	21
5	12
6	4
7	2
	90

Procedure:

A. 1. Compute: -

a. mean

b. standard deviation

c. standard error of the mean

2. determine whether the standard error of the mean is within ± 10% of the mean with a probability of 95:100.

* * * * * * * * * *

Data:

B. A sample of 36 items from a normal population had the following statistics: -

$\bar{X} = 12.4$

$s = \pm 7.6$

Procedure:

B. 1. Determine the standard error of the mean, $s_{\bar{x}}$, at the 5% probability level.

2. If $s_{\bar{x}}$ is not within ± 10% of the sample mean, determine how many sampling units must be measured in order to meet the specification with a probability of 95:100.

3. Using the original statistics, compute the range within which the true mean of the population will lie at the 5% and 1% probability levels.

4. If we postulate the hypothesis that $\bar{X} = 7$, would we accept or reject the hypothesis at the 5% level? At the 1% level? Show the necessary calculations.

Exercise 7

STANDARD ERROR OF A DIFFERENCE

Object:

To determine if the difference between two treatments is significant.

Materials needed:

Column paper
Desk calculator

Data:

In 1960, an experiment was conducted to determine the effect of number of hours of artificial light on the height growth of red pine seedlings. All other growing conditions were kept as uniform as possible. At the end of an 8-week period, height was measured and recorded for 10 seedlings selected at random from each of two treatments. The treatments were (A) 17.5 hours and (B) 7.5 hours of artificial light. The data for the 10 seedlings in each group were: -

Seedling No.	Height growth	
	17.5 hours	7.5 hours
	mm.	mm.
1	25	14
2	28	18
3	29	16
4	24	13
5	22	15
6	26	19
7	25	12
8	29	17
9	27	14
10	23	16

Procedure:

1. In order to calculate the standard error of the difference between the two mean, $s_{\bar{d}}$ you must compute the following statistics for each sample: -

   a. mean

   b. standard deviation

   c. standard error of the mean.

2. Compute the standard error of the difference between the two means from: -

$$s_{\bar{d}} = \sqrt{\frac{2s^2}{n}} \qquad \text{(Formula 6.9)}$$

where $$s^2 = \frac{\Sigma x^2}{2(n-1)} \qquad \text{(Formula 6.8)}$$

3. Compute the value of 't' by Formula 6.7 and determine whether the difference between the two sample means is significant or not. Check your calculations by using Formula 6.10.

Exercise 8

SAMPLING TECHNIQUES

Object:

To obtain random and systematic samples of a forested area and to compare the results statistically.

Materials needed:

Table of random numbers
Column paper
Desk calculator

Data:

Use Figure 7.4 as the basic data (Chapter 7 - Sampling Techniques).

Procedure:

A. 1. Select a 5% random sample from the 400-acre tract using the table of random numbers to select the sampling units. You may use either a row-column combination or number the sampling units from 1 to 400 inclusive.

2. Compute the mean, standard deviation and standard error of your sample.

3. Set the 5% fiducial limits on the population mean value.

4. Compute the volume estimate for the 400-acre tract.

B. 1. Obtain a 10% stratified random sample from Figure 7.4 by using the following number of sampling units by strata: -

Block	No. of sampling units
A	8
B	3
C	14
D	13
E	2

2a. Compute the mean, standard deviation and standard error of the mean for each block.

b. prepare a volume estimate by blocks using $\bar{X} \pm t_{.05} s_{\bar{x}}$ for each estimate.

C. 1. Obtain a 5% systematic sample of the 400-acre tract (Figure 7.4) by using a random start for the row and column as outlined in Chapter 7.

D. Set up appropriate headings for a table to compare all the results of your sampling procedures.

Exercise 9

LINEAR REGRESSION

Object:

To use the least squares method of computing a linear regression equation and to determine the effectiveness of the straight line relationship.

Materials needed:

1 sheet 10 x 10 graph paper
Column paper
Desk calculator

Data:

The data below were extracted from "Growth in Well-stocked Natural Oak Stands in Missouri". Mo. Agr. Exp. Sta. Res. Bull. 700, 20 pp.

Basal area per acre square feet	Sawtimber volume fbm Int. 1/4" rule
43.0	3575
16.0	785
55.0	3740
60.0	4805
49.0	3725
19.0	1825
25.0	1825
29.0	2155
58.0	3965
63.0	5245
25.0	1905
82.0	7810
52.0	5165
50.0	4330
63.0	7760
72.0	6280

Procedure:

1. Compute the linear regression equation which will fit the data. You may use either the method which uses the two normal equations (Formula 8.2) or that using deviations from mean values (Formula 8.3).

2. a. Compute the standard deviation from regression for the data (Formula 8.5).

   b. Compute the standard error of the regression coefficient (Formula 8.6) and determine whether the regression coefficient is significantly different from $\bar{Y}$ (Formula 8.7).

3. Compute the standard error of the predicted $\hat{Y}$ values using Formula 8.8 for the 5% probability level.

4. Draw a graph showing the original points, the regression line and the limits of $s_{\hat{Y}}$ at the 5% probability level.

Exercise 10

LINEAR CORRELATION

Object:

To calculate the linear regression equation and degree of correlation from a given set of data.

Materials needed:

1 sheet 10 x 10 graph paper
Column paper
Desk calculator

Data:

The measurements given below were obtained by F. W. Taylor, University of Missouri, in which he measured the relationship between ring width (X) and vessel volume (Y) of one yellow poplar tree.

Ring width inches	Vessel volume percent
0.379	31.7
0.332	37.8
0.463	35.0
0.229	41.4
0.212	40.8
0.142	37.1
0.035	55.0
0.065	47.0
0.037	52.6
0.047	57.7
0.020	57.3

Procedure:

1. Calculate the linear regression equation using Formula 8.3 and the coefficient of correlation using Formula 8.16. The Σx^2, Σy^2 and Σxy terms can be obtained by cumulating on the calculator. For instance: -

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

$$\Sigma xy = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$$

2. Plot the original points and the position of the regression line calculated in (1) above.

3. How much of the variation in the dependent variable is attributable to the independent variable?

Exercise 11

ANALYSIS OF VARIANCE

Object:

To compute the analysis of variance for one-way classification.

Materials needed:

Column paper
Desk calculator

Data:

Four silvicultural treatments were applied to five plots each to determine the effect of thinning on diameter growth at the end of a 10-year period. The data are given in the table below.

	Diameter growth in inches				
Treatment	Plot 1	Plot 2	Plot 3	Plot 4	Plot 5
1	1.0	0.8	1.5	0.9	1.3
2	2.5	1.7	3.0	2.2	2.6
3	3.4	3.7	2.8	3.6	4.0
4	4.3	4.0	3.9	4.7	4.6

Procedure:

1. Compute the analysis of variance for the data and determine whether the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ should be accepted or rejected.

2. If the null hypothesis is rejected, determine which of the treatments is significantly different from the others by using Snedecor's Q-test.

Exercise 12

ANALYSIS OF VARIANCE (FACTORIAL)

Object:

To compute the analysis of variance for a 2 x 2 factorial.

Materials needed:

Column paper
Desk calculator

Data:

A 2 x 2 factorial consists of two treatments, each applied at two levels to a number of blocks. The data given below is typical of a 2 x 2 factorial: -

Treatment A	Treatment B					
	B_1			B_2		
	Blocks			Blocks		
	1	2	3	1	2	3
A_1	46	40	42	53	58	52
A_2	38	41	36	48	44	43

Procedure:

Follow the procedures outlined in Chapter 9 for the analysis of variance, obtaining sums of squares and mean squares for: -

(a) total (sum of squares only)

(b) treatments combined

(c) treatment A

(d) treatment B

(e) interaction AB

(f) error

and determine the significance of each mean square. Present the results in an analysis of variance table.

FORMULAS

FORMULAS

3.1 Mean, ungrouped data:

$$\bar{X} = \frac{\Sigma X}{N} \text{ or } \frac{\Sigma X}{\Sigma f}$$

3.2 Mean, grouped data:

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

3.3 Mean, using assumed mean:

$$\bar{X} = \bar{X}_A + \frac{\Sigma fx'}{\Sigma f}$$

3.4 Median:

$$X_{me} = L_{me} + \frac{i}{f}(C)$$

3.5 Median:

$$X_{me} = L_{me} + \left(\frac{\frac{N}{2} - \Sigma f_{cum}}{f_{me}}\right)C$$

3.6 Mode:

$$X_{mo} = L_{mo} + \left(\frac{f_H}{f_H + f_L}\right)C$$

4.1 Normal curve:

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

5.1 Standard deviation ungrouped data:

$$s = \sqrt{\frac{\Sigma x^2}{N - 1}}$$

5.2 Standard deviation grouped data:

$$s = \sqrt{\frac{\Sigma fx^2}{\Sigma f - 1}}$$

5.3 Standard deviation, from using assumed mean:

$$s = \sqrt{\frac{\Sigma fx'^2 - \frac{(\Sigma fx')^2}{\Sigma f}}{\Sigma f - 1}}$$

5.4 Normal deviate:

$$z = \frac{X - \bar{X}}{s} \text{ or } \frac{x}{s}$$

5.5 Coefficient of variation:

$$V = \frac{s}{\bar{X}}(100)$$

6.1 Standard error of the mean:

$$s_{\bar{x}} = \frac{s}{\sqrt{\Sigma f}} \text{ or } \frac{s}{\sqrt{N}}$$

6.2 Confidence interval:

$$\bar{X} - s_{\bar{x}} < \mu < \bar{X} + s_{\bar{x}}$$

6.3 t:

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

6.4 5% confidence interval:

$$\bar{X} - t_{.05}s_{\bar{x}} < \mu < \bar{X} + t_{.05}s_{\bar{x}}$$

6.5 Determining N for $s_{\bar{x}}$ to be within 10% of $\bar{X}$ with probability of 95:100.

$$\sqrt{N} = \frac{t_{.05}s}{s_{\bar{x}}}$$

6.6 Null hypothesis:

$$H_0: \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$

6.7 t-test for significance of difference between two means:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

6.8 Variance for two groups of equal size:

$$s^2 = \frac{\Sigma x^2}{2(n - 1)}$$

6.9 Standard error of a difference between two means, equal numbers in each:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{2s^2}{n}}$$

6.10 t-test for two equal-sized groups:

$$t = \bar{X}_1 - \bar{X}_2 \sqrt{\frac{n(n-1)}{\Sigma x^2}}$$

6.11 Standard error of a difference between two means, equal numbers in each:

$$s_{\bar{x}_1 - \bar{x}_1} = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2}$$

6.12 Standard error of a difference for two groups of unequal size:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

6.12 t-test for two groups of unequal size:

$$t = \bar{X}_1 - \bar{X}_2 \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 - n_2)\,\Sigma x^2}}$$

8.1 Straight line formula:

$$\hat{Y} = a + bX$$

8.2 Normal equations for solving linear regression:

I. $\Sigma Y = Na + b\Sigma X$

II. $\Sigma XY = a\Sigma X + b\Sigma X^2$

8.3 Alternate method of solving linear equation from deviations:

$$\hat{Y} - \bar{Y} = b(X - \bar{X})$$

8.4 Sum of squares of residuals:

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

8.5 Standard deviation from linear regression:

$$s_{yx} = \sqrt{\frac{\Sigma d_{yx}^2}{N - 2}}$$

8.6 Standard error of the regression coefficient:

$$s_b = \frac{s_{yx}}{\sqrt{\Sigma x^2}}$$

8.7 t-test for significance of regression coefficient:

$$t = \frac{b}{s_b}$$

8.8 Standard error of $\hat{Y}$ (X not subject to sampling error):

$$s_{\hat{Y}} = s_{yx} \sqrt{\frac{1}{N} + \frac{(X - \bar{X})^2}{\Sigma x^2}}$$

8.9 Standard error of $\hat{Y}$ (X subject to sampling error)

$$s_{\hat{Y}} = s_{yx} \sqrt{1 + \frac{1}{N} + \frac{(X - \bar{X})^2}{\Sigma x^2}}$$

8.10 Formula for second degree curve, parabolic formula:

$$\hat{Y} = a + bX + cX^2$$

8.11 Normal equations for solving second degree curve:

I. $\Sigma Y = Na + b\Sigma X + c\Sigma X^2$

II. $\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$

III. $\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$

8.12 Standard deviation from curvilinear regression (second degree):

$$s_{yx} = \sqrt{\frac{\Sigma d_{yx}^2}{N - 3}}$$

8.13 F-ratio:

$$F = \frac{\text{larger mean square}}{\text{smaller mean square}}$$

8.14 Parabola, general formula:

$$Y = aX^b$$

8.15 Exponential curve, general formula:

$$Y = ab^X$$

8.16 Coefficient of correlation:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

8.17 Relation between r and regression coefficients:

$$r^2 = b_{yx} \cdot b_{xy}$$

8.18 t-test for significance of r:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

PROBLEMS

PROBLEMS

Graphic presentation of data

1.1 In the table below are given the total height in feet and dbh in inches for a number of trees -- all species of oak -- taken in the vicinity of University Forest in Butler County, Missouri.

a. Group the heights into 1" classes of dbh and plot the average points on a sheet of 10 x 10 graph paper.

b. Draw a balanced free-hand curve to satisfy the plotted points.

c. Tabulate the total height for the inch classes of dbh on a separate sheet of column paper.

dbh	Total Height	dbh	Total Height
Inches	Feet	Inches	Feet
3.6	29	7.1	47
3.9	29	7.5	49
3.6	38	7.3	49
2.0	18	7.4	39
5.5	34	8.1	50
4.5	34	8.1	50
4.3	36	8.1	38
2.9	15	7.1	46
3.5	29	7.2	43
4.9	46	7.9	40
4.4	35	9.5	45
5.5	43	8.7	46
5.0	35	9.5	47
6.0	49	8.1	41
5.6	38	8.2	54
5.7	45	7.8	47
4.7	35	7.8	51
4.9	36	10.5	43
7.1	48	9.2	56
6.9	49	8.7	49
7.3	45	10.8	53
6.0	49	11.5	55
6.2	51	10.6	55
6.0	47	11.7	47
6.2	50	12.1	58
6.4	44	11.8	42
7.2	45	12.0	48
7.1	50	11.1	56
		10.6	51
		13.8	48

1.2 Draw a balanced curve to show the relationship between dbh and height for the following black oaks measured in Wayne County, Missouri: -

dbh	f	Total Height
4.8	7	37
5.9	12	42
7.0	20	52
8.1	17	55
9.1	15	62
10.0	7	65
10.9	5	64
12.2	4	70
13.3	2	69
13.9	2	73

1.3 Use the data below to draw a freehand curve: -

X	f	Y
3.6	4	18
4.8	15	26
5.9	24	33
6.9	32	40
8.2	18	45
9.1	7	50
10.2	5	52

Class intervals

2.1 If the height of a tree was measured and found to be 24.7 feet, what height class would you put it into: -

a. using one-foot height classes

b. using two-foot height classes

What are the limits of these classes?

2.2 If you were using 10-foot height classes, in what class would the tree in Problem 2.1 be placed? What are the limits of this class?

2.3 Group the items listed below in 2-foot and 5-foot classes and show the frequency in each class for each grouping: -

0.6, 0.9, 1.0, 1.1, 1.4, 1.4, 1.6, 1.8, 1.8, 1.9, 2.2, 2.4, 2.4, 2.4, 2.4,
2.5, 2.7, 2.7, 2.8, 3.0, 3.0, 3.0, 3.4, 3.5, 3.5, 3.6, 3.8, 4.1, 4.1, 4.4,
4.5, 4.7, 4.7, 4.9, 5.0, 5.5, 5.8, 5.9, 6.3, 6.4, 6.6, 6.8, 7.2, 7.4, 7.7,
7.7, 7.9, 8.2, 8.3, 8.3, 8.6, 8.7, 8.9, 9.2, 9.4, 9.6, 9.8, 10.2, 10.4,
10.8, 11.3, 11.7, 12.2, 12.8, 13.4, 13.6

Measures of central tendency

3.1 Calculate and plot the mean, median and mode of each distribution in Problem 4.1, page 134 Show the calculations of each statistic.

3.2 Find the median in the following distribution: -

X	f
4	3
6	12
8	15
10	6
12	4

3.3 Why is the median not a good indicator of normality when taken by itself?

3.4 Which curve would be skewed to a lesser degree: -

a. $\bar{X} = 32$ Me = 27 Mo = 24

b. $\bar{X} = 32$ Me = 30 Mo = 29

Are both these curves positively, or both negatively, or one positively and the other negatively, skewed? Why?

3.5 Define: mean, median, mode, kutosis, skewness, plati-kurtosis.

3.6 What type of forest stand would have a diameter distribution resembling a normal curve? Explain.

3.7 If the value of $\bar{X} = 4.12$ and Mo = 4.45, where would you expect the median to be located?

3.8 Calculate the mean, median and mode of the distribution shown in Exercise 5, page 117.

3.9 If you were given two figures, one being the mean and the other the median salary among a group of factory workers, which one would be greater in value? Why?

3.10 Calculate the value of the mode in the distribution on page 118, Exercise 6.

3.11 Determine the median for the distribution on page 118, Exercise 6.

3.12 Find the mean, median and mode of: -

X	f
0	1
1	4
2	16
3	32
4	14
5	3

3.13 Determine the average height of these items without grouping them into classes: -

24	24	22	18	16	34	20
22	20	22	24	26	28	26
30	32	34	24	28		

3.14 If the figures in Problem 3.13 were total height in feet, group them into two-foot height classes and obtain the average height by both methods.

3.15 Obtain the mean of this distribution: -

X	f
8	5
10	6
12	3
14	2
16	1

3.16 Use an assumed mean and calculate the true mean of the items in Problem 3.15.

3.17 Obtain the mean of the following enumeration data by using the assumed mean method: -

dbh	f
8	12
10	8
12	10
14	14
16	10
18	6
20	6
22	4

Check your answer by obtaining the mean by the regular method.

3.18 The formula for obtaining the true mean by using an assumed mean is: -

$$\bar{X} = \bar{X}_A + \frac{\Sigma fx'}{\Sigma f}$$

If the value of the assumed mean is larger than the true mean, the correction factor must be subtracted. Why is the formula not: -

$$\bar{X} = \bar{X}_A \pm \frac{\Sigma fx'}{\Sigma f} \; ?$$

3.19 Using the 16-inch class as the assumed mean in the example in Table 5.1, calculate the true mean for each distribution.

Populations, samples and more on frequency distributions

4.1 Plot the following distributions on separate sheets of graph paper: -

A		B		C	
X	f	X	f	X	f
0	0	0	0	0	0
1	5	1	10	1	0
2	17	2	40	2	5
3	45	3	105	3	15
4	112	4	140	4	30
5	140	5	120	5	55
6	115	6	80	6	100
7	45	7	55	7	130
8	18	8	30	8	100
9	5	9	15	9	30
10	0	10	5	10	5

Which of the three curves represents a normal curve? Which is positively skewed? Which negatively skewed?

4.2 The list below is the result of tossing 10 coins at a time and recording the numbers of heads turned up.

Number of heads turned up in 10 coins tossed at one time	f
0	0
1	7
2	43
3	123
4	210
5	245
6	215
7	116
8	34
9	7
10	0
	1000

a. plot the frequency distribution to show the data follow a normal curve pattern

b. plot the cumulative frequency polygon

c. calculate the mean, median and mode of the frequency distribution.

4.3 The following frequency distribution was obtained by tossing 10 coins at a time for a total frequency of 250.

Number of heads turned up in 10 coins tossed at one time	f
0	0
1	2
2	10
3	22
4	51
5	57
6	63
7	32
8	12
9	1
10	0
	250

a. plot the frequency distribution to show the trend of the relationship

b. plot the cumulative frequency polygon

c. calculate the mean, median and mode of the distribution.

Measures of dispersion

5.1 Determine the standard deviation for the sample of volumes, all figures in cords per acre: - 14.0, 8.5, 9.5, 11.0, 16.0, 14.5, 15.5, 10.0, 7.0, 5.0.

5.2 What is the standard deviation in Problem 5.1 expressed as a per cent of the mean volume? What is the total range of volumes?

5.3 Using the appropriate factor from Table 5.5, what is the estimated standard deviation in Problem 5.1 based on the range of values? How many items in a sample would give as accurate an estimate of standard deviation, using the range, as the calculated standard deviation?

5.4 A sample of 15 items has a range in values from the highest to the lowest of 4.6 inches. What would you expect the standard deviation to be? How many sampling units should be measured to obtain the same accuracy as by calculating s?

5.5 The following seedling heights, in inches, were measured in a forest nursery: - 16, 14, 17, 12, 18, 14, 15, 8, 13, 9, 17, 13, 16, 18, 16, 20. Estimate s from the range of items. Calculate s and compare the result with the estimated s.

5.6 Obtain the mean volume and standard deviation for this sample: -

X	f
1	5
2	15
3	30
4	15
5	5

5.7 In samples from normal populations, the standard deviation is roughly one quarter of the range (for samples of at least 25 items). A sample of 25 oven-dry weights of seedlings at the age of 6 weeks ranged from 106 to 143 grams.

a. What would you expect the standard deviation of this sample to be?

b. What are the limits of seedling weight between $\bar{X} \pm 2s$ assuming that the mean weight was 124 grams?

5.8 Calculate the standard deviation by (1) formula using grouped data and (2) formula using grouped data but with an assumed mean for: -

X	f	X	f	X	f
10	2	15	1	20	1
11	3	16	1	21	-
12	5	17	-	22	-
13	2	18	1	23	-
14	2	19	-	24	1

5.9 What are the fiducial limits of the sample in Problem 5.8 for: -

a. $\bar{X} \pm s$

b. $\bar{X} \pm 2s$

c. $\bar{X} \pm 3s$

5.10 Draw the frequency curve for the distribution in Problem 5.8 and state what type of curve it represents.

5.11 Would the distribution in Table 5.4, page 29 have a larger or smaller coefficient of variation than one which has a mean of 12.0 inches and a standard deviation of $\pm$ 4.0 inches? By how much?

5.12 A sample of 10 trees measured for height in an even-aged stand gave these results: - 60, 63, 57, 62, 63, 66, 42, 61, 67, 59. Find $\bar{X}$, s, and coefficient of variation for this sample.

5.13 Compare the variability of these two sets of measurements. Use an assumed mean to determine standard deviation.

a. 44, 32, 28, 48, 50, 30, 36, 40

b. 37, 31, 40, 41, 35, 33, 30, 35

5.14 Define: standard deviation: coefficient of variation. Tell how the two are related.

5.15 A group of seedlings was measured for height, the mean height being 63.84 inches with a standard deviation of ± 13.08 inches. What are the limits of height included in 95% of the sample? Calculate the coefficient of variation.

5.16 Given heights of 72, 53, 69, 82, 43, 48, 57, 62, 79, and 73 feet, compute the relative dispersion without calculating the standard deviation but by using the table of s/range on page 30.

5.17 Is it true that if s_1 is greater than s_2 by 20%, then V_1 is greater than V_2 by 20%? Why or why not?

5.18 The relative dispersion of heights in a pine stand 80 years old is 25%. Would you think the relative dispersion of a pine stand 20 years old on the same type of soil and same aspect would be larger or smaller? Why?

5.19 What per cent of the total area under a normal curve lies between 1.0 and 1.5 standard units? Consult Table A.2 in the Appendix.

5.20 A sample has a mean diameter of 10.0 inches and a standard deviation of ± 2.0 inches. What per cent of the trees is larger than 15 inches? What per cent lies between 4.0 and 6.0 inches? What per cent is less than 4.0 inches?

5.21 If a sample had a mean value of 12.0 and a standard deviation of ± 4.0, what is the probability of an item having a value of 18.0 occurring?

5.22 A sample of oven-dry weights has a mean of 16 grams and s = ± 6 grams. What is the probability of a weight of 30 grams appearing in a sample taken from the population of weights?

5.23 What is the probability of an item having a normal deviate of -1.7 occurring?

5.24 ± 1 normal deviate is usually taken, for practical purposes, to have a probability of 66%. What is the exact probability? for ± 2 normal deviates? for ± 2.5? for ± 3?

5.25 Which items in the following sample are outside the limits of ± 1 normal deviate: - 42, 44, 54, 62, 47, 49, 37, 53, 40, 52? Are any outside the range of ± 2 normal deviates?

5.26 If a person gave you a 9:1 on a bet, what is the probability that he will lose the bet?

5.27 A horse is entered in a race and the odds posted on him are 50:1. According to the bookies, what chance has he of winning the race?

5.28 A sample of trees has an average diameter of 20.0 inches and the standard deviation of the sample is ± 2.0 inches.

a. What per cent of the trees in the stand is: -

(1) outside the range of 16 to 24 inches inclusive?

(2) between 16 and 18 inches?

(3) larger than 24 inches?

b. Would you consider this sample to be from an even-aged or an uneven-aged stand? Why?

5.29 Given the following statistics, determine what per cent of the items is greater than 20 inches and less than 4.0 inches: -

$$\bar{X} = 12.0 \text{ inches}$$

$$s = \pm 4.0 \text{ inches}$$

5.30 A sample of 64 seedlings in a plantation was measured for height at the end of the third growing season. The mean height was 32 inches and the standard deviation was ± 4.0 inches. How many seedlings in a lot of 500 selected at random would be within the range of 24 to 40 inches?

5.31 The following ages were measured in a stand: -

14	16
18	16
12	17
13	27
15	12

Determine whether any of the measurements are abnormal. If you reject any measurements, determine the final standard deviation.

5.32 If $\frac{X - \bar{X}}{s} > 2$, should the item be rejected as being abnormal? Explain why or why not.

5.33 If an item has a normal deviate of -2.7, will the re-calculated mean be higher or lower in value? the standard deviation?

5.34 If $\bar{X} = 40$ and $s = \pm 8$, should $X = 63$ be rejected as being abnormal? $X = 18$?

5.35 Frequently, in determining the age of a stand, one or two trees are suspected of belonging to a previous stand because they are considerably older than some others. How would you determine whether they should be included in the sample?

5.36 Illustrate the theory of rejection of abnormal data by drawing a normal curve and setting upper and lower limits of rejection.

5.37 Is an item having a large deviation from the mean $(X - \bar{X})$ more likely to be rejected from a sample having a low standard deviation or from one having a high standard deviation?

5.38 What actually determines the upper and lower limits of the rejection?

5.39 In a sample with $\bar{X} = 35$ and $s = \pm 5$, what is the probability of an item less than $X = 22$ occurring? Would this item be rejected as being abnormal? What is the coefficient of variation in this sample?

Standard error

6.1 Calculate the mean, standard deviation and standard error for the measurements listed below: -

Tree No.	Total Height Feet	Tree No.	Total Height Feet
1	55	9	38
2	45	10	51
3	40	11	60
4	58	12	58
5	62	13	53
6	45	14	45
7	47	15	41
8	50	16	52

a. Was the number of trees sufficient to bring the standard error within ± 10% of the mean height?

b. If the number of trees was sufficient, how many trees could you eliminate from the sample so that the standard error was exactly ± 10% of the mean height?

c. If the standard deviation had been ± 15 feet, how many trees would have to be in the sample to bring the standard error within ± 10% of the mean height 95 times out of 100?

6.2 A number of one-fifth acre plots was laid out in an area and it was necessary that the accuracy of the sample be within ± 10% of the mean with a probability of 2:1. The results of the sample are: -

Number of plots	10
Mean volume per acre	3000 board feet
Standard deviation	± 1200 board feet

a. Was the objective of accuracy obtained? Show your calculations.

b. If not, how many plots would you have to establish to achieve the ± 10% requirement?

c. What is the coefficient of variation of this sample?

6.3 For this distribution of heights in inches, determine the mean, standard deviation and standard error: -

18, 9, 26, 23, 10, 16, 20, 9, 28

How many items would be required to bring the standard error to exactly ± 10% of the mean with a probability of 99:1? Assume that the value of standard deviation did not change with more plots.

6.4 A sample of 48 trees measured for height gave these statistics:

Mean	50 feet
s	± 10 feet

How many trees must be measured to get the sampling error reduced to ± 10%?

6.5 A random sample of 15 tree diameters resulted in $\bar{X}$ = 12.0 inches and s = ± 5.0 inches. Is the sample of 15 trees sufficient to bring the standard error within ± 5% of the mean 68% of the time? If not, how many trees should be measured?

6 6 As a result of an experiment, it took a sample of 400 white pine seeds 70.5 days to germinate and the standard deviation was ± 8.0 days. Estimate the standard error.

6.7 Calculate the diameter limits within which the true mean of the population will lie with a probability of 95 out of 100 for this sample:

14, 18, 6, 12, 10, 7, 19, 13, 21

6.8 A sample of 49 trees gave a mean height of 60 feet and s = ± 14 feet. How many more trees should be measured to obtain a standard error which is 2% of the mean, with a probability of 2 out of 3?

6.9 A sample consisting of diameter measurements of 36 trees was taken in each of two stands. The statistics are: -

Stand 1		Stand 2	
$\bar{X}_1$	= 10.0 inches	$\bar{X}_2$	= 11.2 inches
s_1	= ± 1.8 inches	s_2	= ± 2.4 inches
$s_{\bar{x}_1}$	= ± 0.3 inches	$s_{\bar{x}_2}$	= ± 0.4 inches
n	= 36	n	= 36

a. Determine whether the difference between the means is significant.

b. Determine the probability (as a percentage) of the difference occurring by chance.

6.10 In the example shown, determine whether the two samples are from the same population or from different ones: -

Sample 1		Sample 2	
$\bar{X}$	= 6.90 inches	$\bar{X}$	= 4.6 inches
s	= ± 2.0 inches	s	= ± 2.0 inches
$s_{\bar{x}}$	= ± 0.5 inches	$s_{\bar{x}}$	= ± 0.5 inches
n	= 16	n	= 16

6.11 If you set up the hypothesis that $\bar{X}_1 - \bar{X}_2 = 0$, what are the conditions under which you would reject the hypothesis at the 5% level of significance?

6.12 If the difference between the means of two samples was greater than $3s_{\bar{d}}$, what is the probability that the two samples came from the same population?

6.13 $\bar{X}_1 - \bar{X}_2 = 2$. State the null hypothesis and draw a valid conclusion as to the population from which the two samples were drawn.

6.14 Describe how you could determine experimentally whether a particular type of fertilizer increases growth rate of pine seedlings in a forest nursery. Briefly write out the steps you would follow in carrying out the experiment and analyzing the results.

6.15 Two different treatments were applied to plots in the same stand; one treatment applied to 49 plots consisted of thinning to a basal area of 70 square feet per acre, while the second treatment was a thinning to 90 square feet per acre applied to 36 plots. At the end of 10 years, average growth in cubic feet per acre per year was measured. These are the results: -

Treatment	Mean Annual Growth in Cubic Feet	s
1	60	± 5 cu. ft.
2	62	± 10 cu. ft.

Determine whether the effect of the treatments was significantly different.

6.16 A fertilizer treatment was applied to one bed of seedlings in a greenhouse, but not to another. At the end of a certain length of time the heights were recorded in inches, as follows: -

No fertilizer: 14, 22, 20, 15, 26, 19, 28, 22, 16, 26

Fertilizer: 32, 24, 30, 36, 20, 25, 35, 28, 32, 28

Assuming that the soil in both beds was the same type and other conditions were equally controlled for both beds, did the fertilizer cause an increase in relative dispersion?

6.17 Show whether you would accept or reject the null hypothesis for the information given on these two samples: -

Sample 1	Sample 2
$\bar{X} = 12.0$	$\bar{X} = 10.7$
$s = \pm 3.1$	$s = \pm 2.5$
$n = 25$	$n = 25$

Sampling techniques

7.1 a. Use the table of random numbers in the Appendix to obtain ten random samples of 10% each from the volumes in Figure 7.3, page 53. Compute the average volume for each sample.

b. Use the factor of 0.231 (s/range with n = 40) and calculate the estimated standard deviation for each sample.

c. Compute the standard error for each sample and determine whether the population mean (18.50 hundreds of board feet) lies within the mean volume $\pm s_{\bar{x}}$. How many of the ten samples were within this figure?

7.2 Take a 10% random sample from each of the stratified areas shown in Figure 7.4, page 55, on the basis of area alone. The number of plots to take is given on page 57. Calculate the same statistics as in the first question and determine whether the sub-population means lie within the sample mean $\pm s_{\bar{x}}$. The sub-population means are: -

Stratum	Mean Volume (00's fbm)
A	17.89
B	7.00
C	24.57
D	40.79
E	5.50

7.3 Take a 5% systematic sample from the volumes in Figure 7.4, page 55, and calculate a total volume estimate for the 400 acres. Knowing the total volume is 7399 hundreds of board feet, is this population total included in the sample mean $\pm s_{\bar{x}}$?

8.1 Explain how a formula for a parabolic type of curve, such as $\hat{Y} = X^2$, would plot as a straight line on double log paper.

8.2 The following information is given as a result of a study of the relation between one variable and another: -

$$N = 25 \qquad \Sigma X = 25 \qquad \Sigma XY = 85$$

$$\Sigma Y = 15 \qquad \Sigma X^2 = 135$$

Solve for "a" and "b" and write the equation for the straight line.

8.3 Find the values for Y when X = 8 in this formula: -

$$\hat{Y} = 1.865 + 0.818X + 0.024X^2$$

8.4 Determine the regression equation by the two methods described in the text for the following data: -

Board Feet 1951	Board Feet 1956
500	600
1000	1200
2500	3200
4000	5000
5000	5500

Verify that b = 1.136. Plot three points and draw the regression line.

8.5 A study of acorn production for scarlet oak as it is influenced by crown diameter resulted in a distribution of points which has a second degree curve trend. The totals for 20 trees are: -

$$\Sigma X = 480$$

$$\Sigma Y = 6{,}508$$

$$\Sigma XY = 198{,}305$$

$$\Sigma X^2 = 13{,}100$$

$$\Sigma X^2Y = 6{,}389{,}675$$

$$\Sigma X^3 = 391{,}500$$

$$\Sigma X^4 = 12{,}447{,}500$$

in which X = crown diameter in feet

Y = total numbers of acorns produced

Verify that $\hat{Y} = 35.56 - 6.4X + 0.677X^2$ and plot the curve on 10 x 10 graph paper.

8.6 If N = 20, $\Sigma X = 360$, $\Sigma Y = 400$, $\Sigma XY = 2880$ and $\Sigma X^2 = 5400$, solve for "a" and "b" and write the equation for the straight line which satisfies these conditions.

Find the value of $\hat{Y}$ when X = 27.

8.7 Transform the following equations into logarithmic form:

a. $Y = \frac{3}{X^2}$

b. $Y = \frac{120}{4}(3^{x^3})$

c. $Y = \left(\frac{a}{b}\right)^{3x}$

d. $Y = 6\sqrt{X^3}$

e. $Y = (X^4)(A^3)$

8.8 If the slope of a straight line is -1/3, and the Y-intercept is -2/3, write the equation of the straight line so that the coefficients of X and Y are whole numbers.

8.9 Write the normal equations for the solution of the least squares method of fitting a straight line. Outline the steps necessary for the solution of the "a' and "b" terms.

8.10 Calculate the value of Y when X = 2 for these equations: -

a. $Y = .60 - 1.8X$

b. $Y^2 = 10.24 + 4X^2$

c. $1.821Y - 3.452X - 15.324 = 0$

Repeat for X = 0; X = 4; X = 10.

8.11 Given N = 20, $\Sigma X = 200$, $\Sigma Y = 300$, $\Sigma XY = 3400$, and $\Sigma X^2 = 6800$. Solve for "a" and "b" and write the straight line equation.

8.12 Substitute values of X = 2, 3, 4, and 6 in the following equations and plot the value of Y on X on single log or double log paper as indicated by the type of formula: -

a. $Y = X^2$

b. $Y = 3(4)^x$

c. $Y = -4.0 - 3X - X^2$

d. $Y = 10^x$

8.13 a. Compute the second degree equation which will fit the following:

ΣX	=	202
ΣY	=	110
ΣXY	=	7,520
ΣX^2	=	8,640
ΣX^2Y	=	214,600
ΣX^3	=	322,560
ΣX^4	=	10,244,600
N	=	16

b. Plot the curve for sufficient values of X to demonstrate the curvilinear relation.

Correlation

8.14 Determine r for the following combination of values, X = volume in board feet and Y = basal area per acre for white oak stands: -

X	Y
46	38
8	4
59	53
2	1
69	48
39	29
47	43
52	42
62	68
53	49
28	19
20	13
38	24
24	18
15	9
58	44

Also show that $r^2 = (b_{yx})(b_{xy})$.

Analysis of variance

9.1 Determine the analysis of variance for the following data consisting of 3 treatments with 6 replications each: -

	Treatment		
Replications	A	B	C
1	42	31	34
2	36	29	30
3	31	34	36
4	46	28	38
5	38	26	34
6	40	30	39

Compute the total sum of squares, treatment sum of squares and error sum of squares. Then compute the mean square for treatment and error and determine the F-value for treatments.

9.2 Is there a significant difference between treatments? Is any one treatment significantly different from the other two? (use the Q-test).

9.3 In a randomized block experiment, treatments are assigned to plots within blocks at random. The data below represent the results of a randomized block experiment in which four treatments are applied to four plots in each block.

	Block			
Plot	1	2	3	4
1	A 90	D 67	D 73	C 72
2	B 73	C 78	C 74	A 85
3	C 76	A 82	B 78	D 66
4	D 68	B 76	A 89	B 79

a. re-arrange the data so that it shows measurements by treatment and plot. (see Table 9.4, page 93).

b. analyze the variance to show the mean square for (1) treatments (2) blocks (3) error and determine which has a significant F-value.

TABLES

Table A.1 - Randomly Assorted Digits*

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35744	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067
20	75884	12952	84318	95108	72305	64620	91318	89872	45375	85436
21	16777	37116	58550	42958	21460	43910	01175	87894	81378	10620
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337
23	42902	66892	46134	91432	94710	23474	20423	60137	60609	13119
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700
28	70577	42866	24969	61210	76046	67699	42054	12696	93758	03283
29	94522	74358	71659	62038	79643	79169	44741	05437	39038	13163
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567
36	45476	84882	65109	96597	25930	66790	65706	61203	53634	22557
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716

*Reproduced by permission from George W. Snedecor: STATISTICAL METHODS (5th Edition 1956), copyright, Iowa State University Press, Ames, Iowa.

Table A.2 -

Areas Under a Normal Curve Between the Maximum Ordinate and the Ordinate at z

z	Area	z	Area	z	Area	z	Area	z	Area
.00	.00000	.40	.15542	.80	.28814	1.20	.38493	1.60	.44520
.01	.00399	.41	.15910	.81	.29103	1.21	.38686	1.61	.44630
.02	.00798	.42	.16276	.82	.29389	1.22	.38877	1.62	.44738
.03	.01197	.43	.16640	.83	.29673	1.23	.39065	1.62	.44845
.04	.01595	.44	.17003	.84	.29955	1.24	.39251	1.64	.44950
.05	.01994	.45	.17364	.85	30234	1.25	.39435	1.65	.45053
.06	.02392	.46	.17724	.86	.30511	1.26	.39617	1.66	.45154
.07	.02790	.47	.18082	.87	.30785	1.27	.39796	1.67	.45254
.08	.03188	.48	.18439	.88	.31057	1.28	.39973	1.68	.45352
.09	.03586	.49	.18793	.89	.31327	1.29	.40147	1.69	.45449
.10	.03983	.50	.19146	.90	.31594	1.30	.40320	1.70	.45543
.11	.04380	.51	.19497	.91	.31859	1.31	.40490	1.71	.45637
.12	.04776	.52	.19847	.92	.32121	1.32	.40658	1.72	.45728
.13	.05172	.53	.20194	.93	.32381	1.33	.40824	1.73	.45818
.14	.05567	.54	.20540	94	.32639	1.34	.40988	1.74	.45907
.15	.05962	.55	.20884	.95	.32894	1.35	.41149	1.75	.45994
.16	.06356	.56	.21226	.96	.33147	1.36	.41309	1.76	.46080
.17	.06749	.57	.21566	.97	.33398	1.37	.41466	1.77	.46164
.18	.07142	.58	.21904	98	.33646	1.38	.41621	1.78	.46246
.19	.07535	.59	.22240	.99	.33891	1.39	.41774	1.79	.46327
.20	.07926	.60	.22575	1.00	.34134	1.40	.41924	1.80	.46407
.21	.08317	.61	.22907	1.02	.34375	1.41	.42073	1.81	.46485
.22	.08706	.62	.23237	1.02	.34614	1.42	.42220	1.82	.46562
.23	.09095	.63	.23565	1.03	.34850	1.43	.42364	1.83	.46638
.24	.09483	.64	.23891	1.04	.35083	1.44	.42507	1.84	.46712
.25	.09871	.65	.24215	1.05	.35314	1.45	.42647	1.85	.46784
.26	.10257	.66	.24537	1.06	.35543	1.46	.42786	1.86	.46856
.27	.10642	.67	.24857	1.07	.35769	1.47	.42922	1.87	.46926
.28	.11026	.68	.25175	1.08	.35993	1.48	.43056	1.88	.46995
.29	11409	.69	.25490	1.09	.36214	1.49	.43189	1.89	.47062
.30	.11791	.70	.25804	1.10	.36433	1.50	.43319	1.90	.47128
.31	.12172	.71	.26115	1.11	.36650	1 51	.43448	1.91	.47193
.32	.12552	.72	.26424	1.12	.36864	1.52	.43574	1.92	.47257
.33	.12930	.73	.26730	1.13	.37076	1.53	.43699	1.92	.47320
.34	.13307	.74	.27035	1.14	.37286	1.54	.43822	1.94	.47381
.35	.13683	.75	.27337	1.15	.37493	1.55	.43943	1.95	.47441
.36	.14058	.76	.27637	1.16	.37698	1.56	.44962	1.96	.47500
.37	.14431	.77	.27935	1.17	.37900	1.57	.44179	1.97	.47558
.38	.14803	.78	.28230	1.18	38100	1.58	.44295	1.98	.47615
.39	.15173	.79	.28524	1.19	.38298	1.59	.44408	1.99	.47670

Table A.2 - (Continued)

Areas Under a Normal Curve Between the Maximum Ordinate and the Ordinate at z

z	Area	z	Area	z	Area	z	Area	z	Area
2.00	.47725	2.40	.49180	2.80	.49744	3.20	.49931	3.60	.49984
2.01	.47778	2.41	.49202	2.81	.49752	3.21	.49934	3.61	.49985
2.02	.47831	2.42	.49224	2.82	.49760	3.22	.49936	3.62	.49985
2.03	.47882	2.43	.49245	2.83	.49767	3.23	.49938	3.63	.49986
2.04	.47932	2.44	.49266	2.84	.49774	3.24	.49940	3.64	.49986
2.05	.47982	2.45	.49286	2.85	.49781	3.25	.49942	3.65	.49987
2.06	.48030	2.46	.49305	2.86	.49788	3.26	.49944	3.66	.49987
2.07	.48077	2.47	.49324	2.87	.49795	3.27	.49946	3.67	.49988
2.08	.48124	2.48	.49343	2.88	.49801	3.28	.49948	3.68	.49988
2.09	.48169	2.49	.49361	2.89	.49807	3.29	.49950	3.69	.49989
2.10	.48214	2.50	.49379	2.90	.49813	3.30	.49952	3.70	.49989
2.11	.48257	2.51	.49396	2.91	.49819	3.31	.49953	3.71	.49990
2.12	.48300	2.52	.49413	2.92	.49825	3.32	.49955	3.72	.49990
2.13	.48341	2.53	.49430	2.93	.49831	3.33	.49957	3.73	.49990
2.14	.48382	2.54	.49446	2.94	.49836	3.34	.49958	3.74	.49991
2.15	.48422	2.55	.49461	2.95	.49841	3.35	.49960	3.75	.49991
2.16	.48461	2.56	.49477	2.96	.49846	3.36	.49961	3.76	.49992
2.17	.48500	2.57	.49492	2.97	.49851	3.37	.49962	3.77	.49992
2.18	.48537	2.58	.49506	2.98	.49856	3.38	.49964	3.78	.49992
2.19	.48574	2.59	.49520	2.99	.49861	3.39	.49965	3.79	.49992
2.20	.48610	2.60	.49534	3.00	.49865	3.40	.49966	3.80	.49993
2.21	.48645	2.61	.49547	3.01	.49869	3.41	.49968	3.81	.49993
2.22	.48679	2.62	.49560	3.02	.49874	3.42	.49969	3.82	.49993
2.23	.48713	2.63	.49573	3.03	.49878	3.43	.49970	3.83	.49994
2.24	.48745	2.64	.49585	3.04	.49882	3.44	.49971	3.84	.49994
2.25	.48778	2.65	.49598	3.05	.49886	3.45	.49972	3.85	.49994
2.26	.48809	2.66	.49609	3.06	.49889	3.46	.49973	3.86	.49994
2.27	.48840	2.67	.49621	3.07	.49893	3.47	.49974	3.87	.49995
2.28	.48870	2.68	.49632	3.08	.49897	3.48	.49975	3.88	.49995
2.29	.48899	2.69	.49643	3.09	.49900	3.49	.49976	3.89	.49995
2.30	.48928	2.70	.49653	3.10	.49903	3.50	.49977	3.90	.49995
2.31	.48956	2.71	.49664	3.11	.49906	3.51	.49978	3.91	.49995
2.32	.48983	2.72	.49674	3.12	.49910	3.52	.49978	3 92	.49996
2.33	.49010	2.73	.49683	3.13	.49913	3.53	.49979	3.93	.49996
2.34	.49036	2.74	.49693	3.14	.49916	3.54	.49980	3.94	.49996
2.35	.49061	2.75	.49702	3.15	.49918	3.55	.49981	3.95	.49996
2.36	.49086	2.76	.49711	3.16	.49921	3.56	.49981	3.96	.49996
2.37	.49111	2.77	.49720	3.17	.49924	3.57	.49982	3.97	.49996
2.38	.49134	2.78	.49728	3.18	.49926	3.58	.49983	3.98	.49997
2.39	.49158	2.79	.49736	3.19	.49929	3.59	.49983	3.99	.49997

Table A.3 - Table of Squares and Square Roots

N	N^2	$\sqrt{N}$	N	N^2	$\sqrt{N}$	N	N^2	$\sqrt{N}$
1.0	1.00	1.0000	4.0	16.00	2.0000	7.0	49.00	2.6458
1.1	1.21	1.0488	4.1	16.81	2.0248	7.1	50.41	2.6646
1.2	1.44	1.0954	4.2	17.64	2.0494	7.2	51.84	2.6833
1.3	1.69	1.1402	4.3	18.49	2.0736	7.3	53.29	2.7019
1.4	1.96	1.1832	4.4	19.36	2.0976	7.4	54.76	2.7203
1.5	2.25	1.2247	4.5	20.25	2.1213	7.5	56.25	2.7386
1.6	2.56	1.2649	4.6	21.16	2.1448	7.6	57.76	2.7568
1.7	2.89	1.3038	4.7	22.09	2.1679	7.7	59.29	2.7749
1.8	3.24	1.3416	4.8	23.04	2.1909	7.8	60.84	2.7928
1.9	3.61	1.3784	4.9	24.01	2.2136	7.9	62.41	2.8107
2.0	4.00	1.4142	5.0	25.00	2.2361	8.0	64.00	2.8284
2.1	4.41	1.4491	5.1	26.01	2.2583	8.1	65.61	2.8460
2.2	4.84	1.4832	5.2	27.04	2.2804	8.2	67.24	2.8636
2.3	5.29	1.5166	5.3	28.09	2.3022	8.3	68.89	2.8810
2.4	5.76	1.5492	5.4	29.16	2.3238	8.4	70.56	2.8983
2.5	6.25	1.5811	5.5	30.25	2.3452	8.5	72.25	2.9155
2.6	6.76	1.6125	5.6	31.36	2.3664	8.6	73.96	2.9326
2.7	7.29	1.6432	5.7	32.49	2.3875	8.7	75.69	2.9496
2.8	7.84	1.6733	5.8	33.64	2.4083	8.8	77.44	2.9665
2.9	8.41	1.7029	5.9	34.81	2.4290	8.9	79.21	2.9833
3.0	9.00	1.7321	6.0	36.00	2.4495	9.0	81.00	3.0000
3.1	9.61	1.7607	6.1	37.21	2.4698	9.1	82.81	3.0166
3.2	10.24	1.7889	6.2	38.44	2.4900	9.2	84.64	3.0332
3.3	10.89	1.8166	6.3	39.69	2.5100	9.3	86.49	3.0496
3.4	11.56	1.8439	6.4	40.96	2.5298	9.4	88.36	3.0659
3.5	12.25	1.8708	6.5	42.25	2.5495	9.5	90.25	3.0822
3.6	12.96	1.8974	6.6	43.56	2.5690	9.6	92.16	3.0984
3.7	13.69	1.9235	6.7	44.89	2.5884	9.7	94.09	3.1145
3.8	14.44	1.9494	6.8	46.24	2.6077	9.8	96.04	3.1305
3.9	15.21	1.9748	6.9	47.61	2.6268	9.9	98.01	3.1464
						10.0	100.00	3.1623

Table A.4 - Values of F for various degrees of freedom and for 5% and 1% points

5% (Roman Type) and 1% (Bold Face Type) Points for the Distribution of F

f_2	f_1 degrees of freedom (for greater mean square)																								f_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
1	161 4,052	200 4,999	216 5,403	225 5,625	230 5,764	234 5,859	237 5,928	239 5,981	241 6,022	242 6,056	243 6,082	244 6,106	245 6,142	246 6,169	248 6,208	249 6,234	250 6,258	251 6,286	252 6,302	253 6,323	253 6,334	254 6,352	254 6,361	254 6,366	1
2	18.51 98.49	19.00 99.00	19.16 99.17	19.25 99.25	19.30 99.30	19.33 99.33	19.36 99.34	19.37 99.36	19.38 99.38	19.39 99.40	19.40 99.41	19.41 99.42	19.42 99.43	19.43 99.44	19.44 99.45	19.45 99.46	19.46 99.47	19.47 99.48	19.47 99.48	19.48 99.49	19.49 99.49	19.49 99.49	19.50 99.50	19.50 99.50	2
3	10.13 34.12	9.55 30.82	9.28 29.46	9.12 28.71	9.01 28.24	8.94 27.91	8.88 27.67	8.84 27.49	8.81 27.34	8.78 27.23	8.76 27.13	8.74 27.05	8.71 26.92	8.69 26.83	8.66 26.69	8.64 26.60	8.62 26.50	8.60 26.41	8.58 26.35	8.57 26.27	8.56 26.23	8.54 26.18	8.54 26.14	8.53 26.12	3
4	7.71 21.20	6.94 18.00	6.59 16.69	6.39 15.98	6.26 15.52	6.16 15.21	6.09 14.98	6.04 14.80	6.00 14.66	5.96 14.54	5.93 14.45	5.91 14.37	5.87 14.24	5.84 14.15	5.80 14.02	5.77 13.93	5.74 13.83	5.71 13.74	5.70 13.69	5.68 13.61	5.66 13.57	5.65 13.52	5.64 13.48	5.63 13.46	4
5	6.61 16.26	5.79 13.27	5.41 12.06	5.19 11.39	5.05 10.97	4.95 10.67	4.88 10.45	4.82 10.27	4.78 10.15	4.74 10.05	4.70 9.96	4.68 9.89	4.64 9.77	4.60 9.68	4.56 9.55	4.53 9.47	4.50 9.38	4.46 9.29	4.44 9.24	4.42 9.17	4.40 9.13	4.38 9.07	4.37 9.04	4.36 9.02	5
6	5.99 13.74	5.14 10.92	4.76 9.78	4.53 9.15	4.39 8.75	4.28 8.47	4.21 8.26	4.15 8.10	4.10 7.98	4.06 7.87	4.03 7.79	4.00 7.72	3.96 7.60	3.92 7.52	3.87 7.39	3.84 7.31	3.81 7.23	3.77 7.14	3.75 7.09	3.72 7.02	3.71 6.99	3.69 6.94	3.68 6.90	3.67 6.88	6
7	5.59 12.25	4.74 9.55	4.35 8.45	4.12 7.85	3.97 7.46	3.87 7.19	3.79 7.00	3.73 6.84	3.68 6.71	3.63 6.62	3.60 6.54	3.57 6.47	3.52 6.35	3.49 6.27	3.44 6.15	3.41 6.07	3.38 5.98	3.34 5.90	3.32 5.85	3.29 5.78	3.28 5.75	3.25 5.70	3.24 5.67	3.23 5.65	7
8	5.32 11.26	4.46 8.65	4.07 7.59	3.84 7.01	3.69 6.63	3.58 6.37	3.50 6.19	3.44 6.03	3.39 5.91	3.34 5.82	3.31 5.74	3.28 5.67	3.23 5.56	3.20 5.48	3.15 5.36	3.12 5.28	3.08 5.20	3.05 5.11	3.03 5.06	3.00 5.00	2.98 4.96	2.96 4.91	2.94 4.88	2.93 4.86	8
9	5.12 10.56	4.26 8.02	3.86 6.99	3.63 6.42	3.48 6.06	3.37 5.80	3.29 5.62	3.23 5.47	3.18 5.35	3.13 5.26	3.10 5.18	3.07 5.11	3.02 5.00	2.98 4.92	2.93 4.80	2.90 4.73	2.86 4.64	2.82 4.56	2.80 4.51	2.77 4.45	2.76 4.41	2.73 4.36	2.72 4.33	2.71 4.31	9
10	4.96 10.04	4.10 7.56	3.71 6.55	3.48 5.99	3.33 5.64	3.22 5.39	3.14 5.21	3.07 5.06	3.02 4.95	2.97 4.85	2.94 4.78	2.91 4.71	2.86 4.60	2.82 4.52	2.77 4.41	2.74 4.33	2.70 4.25	2.67 4.17	2.64 4.12	2.61 4.05	2.59 4.01	2.56 3.96	2.55 3.93	2.54 3.91	10
11	4.84 9.65	3.98 7.20	3.59 6.22	3.36 5.67	3.20 5.32	3.09 5.07	3.01 4.88	2.95 4.74	2.90 4.63	2.86 4.54	2.82 4.46	2.79 4.40	2.74 4.29	2.70 4.21	2.65 4.10	2.61 4.02	2.57 3.94	2.53 3.86	2.50 3.80	2.47 3.74	2.45 3.70	2.42 3.66	2.41 3.62	2.40 3.60	11
12	4.75 9.33	3.88 6.93	3.49 5.95	3.26 5.41	3.11 5.06	3.00 4.82	2.92 4.65	2.85 4.50	2.80 4.39	2.76 4.30	2.72 4.22	2.69 4.16	2.64 4.05	2.60 3.98	2.54 3.86	2.50 3.78	2.46 3.70	2.42 3.61	2.40 3.56	2.36 3.49	2.35 3.46	2.32 3.41	2.31 3.38	2.30 3.36	12
13	4.67 9.07	3.80 6.70	3.41 5.74	3.18 5.20	3.02 4.86	2.92 4.62	2.84 4.44	2.77 4.30	2.72 4.19	2.67 4.10	2.63 4.02	2.60 3.96	2.55 3.85	2.51 3.78	2.46 3.67	2.42 3.59	2.38 3.51	2.34 3.42	2.32 3.37	2.28 3.30	2.26 3.27	2.24 3.21	2.22 3.18	2.21 3.16	13
14	4.60 8.86	3.74 6.51	3.34 5.56	3.11 5.03	2.96 4.69	2.85 4.46	2.77 4.28	2.70 4.14	2.65 4.03	2.60 3.94	2.56 3.86	2.53 3.80	2.48 3.70	2.44 3.62	2.39 3.51	2.35 3.43	2.31 3.34	2.27 3.26	2.24 3.21	2.21 3.14	2.19 3.11	2.16 3.06	2.14 3.02	2.13 3.00	14
15	4.54 8.68	3.68 6.36	3.29 5.42	3.06 4.89	2.90 4.56	2.79 4.32	2.70 4.14	2.64 4.00	2.59 3.89	2.55 3.80	2.51 3.73	2.48 3.67	2.43 3.56	2.39 3.48	2.33 3.36	2.29 3.29	2.25 3.20	2.21 3.12	2.18 3.07	2.15 3.00	2.12 2.97	2.10 2.92	2.08 2.89	2.07 2.87	15
16	4.49 8.53	3.63 6.23	3.24 5.29	3.01 4.77	2.85 4.44	2.74 4.20	2.66 4.03	2.59 3.89	2.54 3.78	2.49 3.69	2.45 3.61	2.42 3.55	2.37 3.45	2.33 3.37	2.28 3.25	2.24 3.18	2.20 3.10	2.16 3.01	2.13 2.96	2.09 2.89	2.07 2.86	2.04 2.80	2.02 2.77	2.01 2.75	16
17	4.45 8.40	3.59 6.11	3.20 5.18	2.96 4.67	2.81 4.34	2.70 4.10	2.62 3.93	2.55 3.79	2.50 3.68	2.45 3.59	2.41 3.52	2.38 3.45	2.33 3.35	2.29 3.27	2.23 3.16	2.19 3.08	2.15 3.00	2.11 2.92	2.08 2.86	2.04 2.79	2.02 2.76	1.99 2.70	1.97 2.67	1.96 2.65	17
18	4.41 8.28	3.55 6.01	3.16 5.09	2.93 4.58	2.77 4.25	2.66 4.01	2.58 3.85	2.51 3.71	2.46 3.60	2.41 3.51	2.37 3.44	2.34 3.37	2.29 3.27	2.25 3.19	2.19 3.07	2.15 3.00	2.11 2.91	2.07 2.83	2.04 2.78	2.00 2.71	1.98 2.68	1.95 2.62	1.93 2.59	1.92 2.57	18
19	4.38 8.18	3.52 5.93	3.13 5.01	2.90 4.50	2.74 4.17	2.63 3.94	2.55 3.77	2.48 3.63	2.43 3.52	2.38 3.43	2.34 3.36	2.31 3.30	2.26 3.19	2.21 3.12	2.15 3.00	2.11 2.92	2.07 2.84	2.02 2.76	2.00 2.70	1.96 2.63	1.94 2.60	1.91 2.54	1.90 2.51	1.88 2.49	19
20	4.35 8.10	3.49 5.85	3.10 4.94	2.87 4.43	2.71 4.10	2.60 3.87	2.52 3.71	2.45 3.56	2.40 3.45	2.35 3.37	2.31 3.30	2.28 3.23	2.23 3.13	2.18 3.05	2.12 2.94	2.08 2.86	2.04 2.77	1.99 2.69	1.96 2.63	1.92 2.56	1.90 2.53	1.87 2.47	1.85 2.44	1.84 2.42	20
21	4.32 8.02	3.47 5.78	3.07 4.87	2.84 4.37	2.68 4.04	2.57 3.81	2.49 3.65	2.42 3.51	2.37 3.40	2.32 3.31	2.28 3.24	2.25 3.17	2.20 3.07	2.15 2.99	2.09 2.88	2.05 2.80	2.00 2.72	1.96 2.63	1.93 2.58	1.89 2.51	1.87 2.47	1.84 2.42	1.82 2.38	1.81 2.36	21
22	4.30 7.94	3.44 5.72	3.05 4.82	2.82 4.31	2.66 3.99	2.55 3.76	2.47 3.59	2.40 3.45	2.35 3.35	2.30 3.26	2.26 3.18	2.23 3.12	2.18 3.02	2.13 2.94	2.07 2.83	2.03 2.75	1.98 2.67	1.93 2.58	1.91 2.53	1.87 2.46	1.84 2.42	1.81 2.37	1.80 2.33	1.78 2.31	22
23	4.28 7.88	3.42 5.66	3.03 4.76	2.80 4.26	2.64 3.94	2.53 3.71	2.45 3.54	2.38 3.41	2.32 3.30	2.28 3.21	2.24 3.14	2.20 3.07	2.14 2.97	2.10 2.89	2.04 2.78	2.00 2.70	1.96 2.62	1.91 2.53	1.88 2.48	1.84 2.41	1.82 2.37	1.79 2.32	1.77 2.28	1.76 2.26	23
24	4.26 7.82	3.40 5.61	3.01 4.72	2.78 4.22	2.62 3.90	2.51 3.67	2.43 3.50	2.36 3.36	2.30 3.25	2.26 3.17	2.22 3.09	2.18 3.03	2.13 2.93	2.09 2.85	2.02 2.74	1.98 2.66	1.94 2.58	1.89 2.49	1.86 2.44	1.82 2.36	1.80 2.33	1.76 2.27	1.74 2.23	1.73 2.21	24
25	4.24 7.77	3.38 5.57	2.99 4.68	2.76 4.18	2.60 3.86	2.49 3.63	2.41 3.46	2.34 3.32	2.28 3.21	2.24 3.13	2.20 3.05	2.16 2.99	2.11 2.89	2.06 2.81	2.00 2.70	1.96 2.62	1.92 2.54	1.87 2.45	1.84 2.40	1.80 2.32	1.77 2.29	1.74 2.23	1.72 2.19	1.71 2.17	25
30	4.17 7.56	3.32 5.39	2.92 4.51	2.69 4.02	2.53 3.70	2.42 3.47	2.34 3.30	2.27 3.17	2.21 3.06	2.16 2.98	2.12 2.90	2.09 2.84	2.04 2.74	1.99 2.66	1.93 2.55	1.89 2.47	1.84 2.38	1.79 2.29	1.76 2.24	1.72 2.16	1.69 2.13	1.66 2.07	1.64 2.03	1.62 2.01	30
40	4.08 7.31	3.23 5.18	2.84 4.31	2.61 3.83	2.45 3.51	2.34 3.29	2.25 3.12	2.18 2.99	2.12 2.88	2.07 2.80	2.04 2.73	2.00 2.66	1.95 2.56	1.90 2.49	1.84 2.37	1.79 2.29	1.74 2.20	1.69 2.11	1.66 2.05	1.61 1.97	1.59 1.94	1.55 1.88	1.53 1.84	1.51 1.81	40
50	4.03 7.17	3.18 5.06	2.79 4.20	2.56 3.72	2.40 3.41	2.29 3.18	2.20 3.02	2.13 2.88	2.07 2.78	2.02 2.70	1.98 2.62	1.95 2.56	1.90 2.46	1.85 2.39	1.78 2.26	1.74 2.18	1.69 2.10	1.63 2.00	1.60 1.94	1.55 1.86	1.52 1.82	1.48 1.76	1.46 1.71	1.44 1.68	50
60	4.00 7.08	3.15 4.98	2.76 4.13	2.52 3.65	2.37 3.34	2.25 3.12	2.17 2.95	2.10 2.82	2.04 2.72	1.99 2.63	1.95 2.56	1.92 2.50	1.86 2.40	1.81 2.32	1.75 2.20	1.70 2.12	1.65 2.03	1.59 1.93	1.56 1.87	1.50 1.79	1.48 1.74	1.41 1.68	1.41 1.63	1.39 1.60	60
70	3.98 7.01	3.13 4.92	2.74 4.08	2.50 3.60	2.35 3.29	2.23 3.07	2.14 2.91	2.07 2.77	2.01 2.67	1.97 2.59	1.93 2.51	1.89 2.45	1.84 2.35	1.79 2.28	1.72 2.15	1.67 2.07	1.62 1.98	1.56 1.88	1.53 1.82	1.47 1.74	1.45 1.69	1.39 1.62	1.37 1.56	1.35 1.53	70
80	3.96 6.96	3.11 4.88	2.72 4.04	2.48 3.56	2.33 3.25	2.21 3.04	2.12 2.87	2.05 2.74	1.99 2.64	1.95 2.55	1.91 2.48	1.88 2.41	1.82 2.32	1.77 2.24	1.70 2.11	1.65 2.03	1.60 1.94	1.54 1.84	1.51 1.78	1.45 1.70	1.42 1.65	1.38 1.57	1.35 1.52	1.32 1.49	80
100	3.94 6.90	3.09 4.82	2.70 3.98	2.46 3.51	2.30 3.20	2.19 2.99	2.10 2.82	2.03 2.69	1.97 2.59	1.92 2.51	1.88 2.43	1.85 2.36	1.79 2.26	1.75 2.19	1.68 2.06	1.63 1.98	1.57 1.89	1.51 1.79	1.48 1.73	1.42 1.64	1.39 1.59	1.34 1.51	1.30 1.46	1.28 1.43	100
200	3.89 6.76	3.04 4.71	2.65 3.88	2.41 3.41	2.26 3.11	2.14 2.90	2.05 2.73	1.98 2.60	1.92 2.50	1.87 2.41	1.83 2.34	1.80 2.28	1.74 2.17	1.69 2.09	1.62 1.97	1.57 1.88	1.52 1.79	1.45 1.69	1.42 1.62	1.35 1.53	1.32 1.48	1.26 1.39	1.22 1.33	1.19 1.28	200
400	3.86 6.70	3.02 4.66	2.62 3.83	2.39 3.36	2.23 3.06	2.12 2.85	2.03 2.69	1.96 2.55	1.90 2.46	1.85 2.37	1.81 2.29	1.78 2.23	1.72 2.12	1.67 2.04	1.60 1.92	1.54 1.84	1.49 1.74	1.42 1.64	1.38 1.57	1.32 1.47	1.28 1.42	1.22 1.32	1.16 1.24	1.13 1.19	400
1000	3.85 6.66	3.00 4.62	2.61 3.80	2.38 3.34	2.22 3.04	2.10 2.82	2.02 2.66	1.95 2.53	1.89 2.43	1.84 2.34	1.80 2.26	1.76 2.20	1.70 2.09	1.65 2.01	1.58 1.89	1.53 1.81	1.47 1.71	1.41 1.61	1.36 1.54	1.30 1.44	1.26 1.38	1.19 1.28	1.13 1.19	1.08 1.11	1000
∞	3.84 6.64	2.99 4.60	2.60 3.78	2.37 3.32	2.21 3.02	2.09 2.80	2.01 2.64	1.94 2.51	1.88 2.41	1.83 2.32	1.79 2.24	1.75 2.18	1.69 2.07	1.64 1.99	1.57 1.87	1.52 1.79	1.46 1.69	1.40 1.59	1.35 1.52	1.28 1.41	1.24 1.36	1.17 1.25	1.11 1.15	1.00 1.00	∞

REFERENCES

Cochran, W. G. (1963). Sampling Techniques 2nd ed. New York. John Wiley and Sons

_____ and Gertrude M. Cox (1950). Experimental Design. New York, John Wiley and Sons

Coile, T. S. (1952). Soils and the growth of forests. Advances in Agronomy, Vol. IV. 330-398

Edwards, Allen L. (1950). Experimental Design in Psychological Research. Rinehart and Company Inc. New York

Fisher, R. A. (1924). Proceedings of the International Mathematical Congress. Toronto, Ontario

_____(1942). Design Of Experiments. Oliver and Boyd. London

Galton, Francis (1888). Proceedings of the Royal Society of London. 45:135.

Hansen, Morris H., William N. Hurwitz and William G. Madow (1953). Sample Survey Methods and Theory. John Wiley and Sons, New York

Loetsch, F. and K. E. Haller (1964). Forest Inventory, Vol. 1. BLV Verlagsgesellschaft Munchen Basel Wien

Nash, Andrew J. (1959). Growth in Well-stocked Natural Oak Stands in Missouri. Mo. Agr. Exp. Sta. Res. Bull. 700. 20 pp.

Snedecor, George W. (1956). Statistical Methods (5th ed.) Iowa State University Press, Ames, Iowa.

Steel, Robert G. D. and James H. Torrie (1960). Principles and Procedures of Statistics. McGraw-Hill Book Company, New York

"Student" (1908) Biometrika, 6.1.

Zahner, Robert (1957). Field Procedures for Soil-site Classification of Pine Land in Southern Arkansas and North Louisiana. U. S. For. Ser. So. For. Exp. Sta. Occ. Paper 155, 17 pp.